УДК 519.218.82 *DOI 10.55421/1998-7072_2025_28_2_70*

З. З. Мингалиев

МЕТОД АДАПТИВНОГО ПРОГНОЗИРОВАНИЯ ДАННЫХ, ЗАВИСЯЩИХ ОТ ВРЕМЕНИ, НА ОСНОВЕ LSTM-МОДЕЛЕЙ С ИСПОЛЬЗОВАНИЕМ КЛАСТЕРИЗАЦИИ ВРЕМЕННЫХ РЯДОВ

Ключевые слова: временные ряды, параметризация, статистические характеристики, спектральные характеристики, кластеризация, прогнозирование, классификация, анализ данных, обучение моделей.

В статье представлен метод адаптивного прогнозирования временных рядов, зависящих от времени, с использованием параметризации, кластеризации и моделей глубокого обучения. Основной целью исследования является повышение точности предсказания временных рядов путем выделения ключевых характеристик данных и создания специализированных моделей для различных групп данных. Предложенный метод включает несколько этапов. В первую очередь выполняется параметризация временных рядов, в ходе которой исходные данные представляются в виде векторов признаков, описывающих их основные статистические и спектральные характеристики (среднее, дисперсия, количество пиков, коэффициенты ARIMA и другие). Затем осуществляется кластеризация временных рядов с помощью алгоритма DBSCAN, что позволяет разделить данные на группы со схожими характеристиками. Для каждой из полученных групп обучается индивидуальная модель на основе рекуррентных нейронных сетей (LSTM), что позволяет учитывать нелинейные зависимости и временную структуру данных. Метод был протестирован на медицинских данных, полученных с инсулиновых помп, фиксирующих уровень глюкозы в крови. Результаты показали, что предложенный подход значительно повышает точность прогнозирования по сравнению с традиционными методами, такими как LSTM без кластеризации или методы на основе евклидова расстояния. Эксперименты подтвердили, что использование параметризации и кластеризации позволяет улучшить адаптивность моделей и повысить точность прогнозов до 95% и выше. Данный подход может быть полезен в различных областях, включая медицину, финансы и энергетику, где требуется точное прогнозирование временных рядов в условиях высокой изменчивости данных.

Z. Z. Mingaliev

ADAPTIVE FORECASTING METHOD FOR TIME-DEPENDENT DATA BASED ON LSTM MODELS USING TIME SERIES CLUSTERING

Keywords: time series, parameterization, statistical characteristics, spectral characteristics, clustering, forecasting, classification, data analysis, model training.

The article presents a method for adaptive forecasting of time-dependent time series using parameterization, clustering, and deep learning models. The main objective of the study is to improve the accuracy of time series forecasting by identifying key data characteristics and creating specialized models for different groups of data. The proposed method includes several stages. First, time series are parameterized, during which the original data are presented as feature vectors describing their main statistical and spectral characteristics (mean, variance, number of peaks, ARIMA coefficients, etc.). Then, time series are clustered using the DBSCAN algorithm, which allows dividing the data into groups with similar characteristics. For each of the resulting groups, an individual model is trained based on recurrent neural networks (LSTM), which allows taking into account nonlinear dependencies and the temporal structure of the data. The method was tested on medical data obtained from insulin pumps that record blood glucose levels. The results showed that the proposed approach significantly improves forecasting accuracy compared to traditional methods such as LSTM without clustering or Euclidean distance-based methods. Experiments confirmed that the use of parameterization and clustering improves the adaptability of models and increases forecast accuracy to 95% and above. This approach can be useful in various fields, including medicine, finance, and energy, where accurate time series forecasting is required under conditions of high data variability.

Введение

Прогнозирование временных рядов является одной из ключевых задач в аналитике данных. В различных отраслях, таких как финансы, медицина, энергетика и другие, точное прогнозирование временных рядов играет критическую роль в принятии решений. Однако создание универсальных моделей для прогнозирования временных рядов затруднено из-за их высокой вариативности и зависимости от множества факторов – сезонность, тренды, случайные шумы, внезапные аномалии, внешние воздействия (например, изменения экономической ситуации или погодных условий) и т.д. [1].

Основная проблема заключается в том, что временные ряды часто обладают сложной структурой, включающей нелинейности, нестабильность и индивидуальные особенности [2]. Стандартные подходы, такие как ARIMA или линейные модели, оказываются малоэффективными в таких условиях, так как они предполагают стационарность данных и линейность зависимости. Кроме того, данные временных рядов могут быть представлены в гетерогенной форме, включающей различные временные разрешения и параметры, что усложняет их обработку стандартными методами. Например, гетерогенными временными рядами являются данные об уровне сахара крови в зависимости от трех временных последовательностей – предыдущего уровня гликемии, уровня углеводов в крови, и количество активных углеводов.

Дополнительная сложность состоит в необходимости персонализации прогнозов. Например, временные ряды одного типа (например, данные о погоде, продажах или медицинских показателях) могут существенно отличаться в зависимости от контекста или субъекта. Универсальные модели не способны эффективно учитывать такие индивидуальные особенности, что приводит к снижению точности прогнозов.

В рамках данной работы предлагается подход, основанный на двухэтапной стратегии:

1. Исходные временные ряды представлены в виде векторов признаков, характеризующих их структуру. Для этого используется метод параметрического описания временных рядов [3]. Затем векторы кластеризуются с помощью алгоритма DBSCAN для разделения данных на группы со схожими характеристиками.
2. Для каждого выделенного кластера обучаются специализированные нейронные сети на основе архитектуры LSTM, способные учитывать нелинейные зависимости и временную структуру данных. Новый временной ряд классифицируется в один из кластеров, после чего прогноз выполняется с использованием модели, обученной на данных этого кластера.

Таким образом, задача прогнозирования временных рядов решается путем кластеризации данных и применения адаптивных подходов на основе глубокого обучения. Такой подход позволяет:

- Учитывать индивидуальные особенности временных рядов;
- Повышать точность прогнозирования за счет специализированных моделей для каждого кластера;
- Обрабатывать гетерогенные и нелинейные данные, а также эффективно справляться с шумами и пропусками в исходных данных.

Предложенный подход предоставляет новые возможности для анализа временных рядов в условиях высокой сложности и вариативности данных.

Существующие методы кластеризации временных рядов

Кластеризация временных рядов — это процесс группировки данных на основе их схожести, который позволяет анализировать и обрабатывать сложные временные структуры. В отличие от классической кластеризации, разделяющей на группы вектора данных, кластеризация временных рядов должна группировать похожие в том или ином смысле последовательности векторов, развернутые во времени. Для этого используются различные методы, представляющие временную последовательность в виде числа или вектора, не зависящих от времени, и последующей кластеризации полученных векторов. Основная цель кластеризации — выделить группы временных рядов, обладающих схожими характеристиками, что упрощает их последующую обработку и анализ. Таким образом, для преобразования временного ряда в вектор, необходимо определить ключевые характеристики, которые и будут представлять временной ряд в виде, независящем от времени.

Существует множество методов кластеризации временных рядов, которые можно разделить на следующие категории:

Методы на основе расстояния

Эти подходы используют меры расстояния между временными рядами для определения их сходства. Наиболее популярные метрики включают:

- Евклидово расстояние [4];
- Корреляционные метрики: такие как коэффициент Пирсона или Спирмена, которые измеряют степень зависимости между временными рядами [4];
- Динамическое время (DTW, Dynamic Time Warping) – метод, позволяющий выравнивать временные ряды с учетом временных сдвигов [5].

Недостатки:

- Евклидово расстояние не учитывает временные искажения и чувствительно к шуму;
- DTW требует больших вычислительных затрат, особенно для длинных временных рядов, и плохо справляется с высокоразмерными данными;
- Корреляционные метрики не подходят для данных с нелинейными зависимостями;
- Методы на основе расстояния неэффективны для кластеризации гетерогенных данных.

Модели на основе признаков

Этот подход предполагает предварительное преобразование временных рядов в векторы признаков, которые описывают их ключевые характеристики. Признаками могут быть:

- Частотные характеристики, полученные с помощью преобразования Фурье [6] или вейвлет-преобразования [7].
- Автокорреляционные свойства и параметры моделей ARIMA [8].

Недостатки:

- Потенциальная потеря важной информации о временных зависимостях.
- Необходимость ручного выбора и вычисления признаков, что увеличивает сложность работы.
- Невозможность точно описать сложные и нелинейные структуры временных рядов с помощью ограниченного набора признаков.
- Чувствительность к качеству данных: наличие шумов или выбросов может сильно повлиять на результаты кластеризации.

Модели на основе глубокого обучения

Современные подходы используют архитектуры глубокого обучения для извлечения скрытых признаков временных рядов. Среди популярных методов:

- Автоэнкодеры [9];
- Рекуррентные нейронные сети (RNN) [10];
- Генеративно-состязательные сети (GAN) [11];

Недостатки:

- Высокая вычислительная сложность, особенно для длинных временных рядов;
- Требуют больших объемов данных для обучения, что не всегда доступно;
- Чувствительность к гиперпараметрам и архитектуре сети;
- Сложность интерпретации результатов кластеризации.

Методология

Для решения задачи кластеризации временных рядов и создания прогнозных моделей для каждого кластера была разработана методология, которая включает несколько ключевых этапов.

Построение модели

Этап 1

На первом этапе данные подвергаются предобработке. Каждый временной ряд разделяется на блоки фиксированного размера, что позволяет упростить анализ и обработку данных. Пропущенные значения заполняются с помощью линейной интерполяции, что сохраняет временную структуру. Далее данные сглаживаются с использованием скользящего среднего, чтобы уменьшить влияние шума и улучшить выделение характеристик временных рядов [12].

Этап 2

Следующим этапом является более глубокое выделение ключевых параметров временных рядов, которые помогают захватить как общие статистические характеристики, так и более сложные динамические особенности временной серии. Для каждого блока временного ряда рассчитываются несколько типов параметров, которые служат основой для последующего анализа и кластеризации.

Первоначально, для каждого блока временного ряда вычисляются стандартные статистические параметры, такие как минимальное, максимальное, среднее и медианное значения. Эти базовые характеристики дают общее представление о распределении данных в пределах блока. Среднее значение помогает понять общий уровень временного ряда, тогда как медиана является более устойчивым показателем, менее чувствительным к выбросам. Минимум и максимум указывают на экстремальные значения, которые могут быть важными для выявления аномальных состояний в данных.

Для более глубокой оценки вариативности и изменчивости временного ряда рассчитывается дисперсия (1). Этот параметр оценивает степень рассеяния значений вокруг среднего, помогая выявить, насколько данные стабильны или изменчивы.

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad (1)$$

Далее вычисляется количество пиков и впадин в сглаженных данных. Пики и впадины могут быть индикаторами значимых событий в ряду, таких как экстремальные изменения или переходные моменты. Пики часто отражают точки, где данные достигают локального максимума, а впадины — локального минимума. Эти параметры полезны для выявления циклических и периодических явлений в данных, таких как сезонные колебания или регулярные аномалии.

Площадь под графиком временного ряда (2) также является важным параметром, так как она отражает общее количество энергии или объема, представленного данными. Эта величина может быть полезна для понимания масштаба изменений и общей тенденции временного ряда.

$$S = \int X(t)dt \qquad (2)$$

Сумма квадратов производной временного ряда (3) оценивает скорость изменений во временном ряду, давая представление о том, насколько быстро происходят колебания. Это особенно полезно для выявления резких изменений или волатильности в данных, таких как скачки или резкие падения.

$$Sum = \sum_{i=1}^{n}\left(\frac{X_{i+1} + X_i}{\Delta t}\right)^2 \qquad (3)$$

Показатель тренда, который рассчитывается с помощью линейной регрессии, представляет собой наклон прямой, аппроксимирующей данные. Он помогает оценить, существует ли устойчивое направление в изменениях временного ряда (например, рост или падение) и какова его степень.

Наконец, количество пересечений временного ряда с определенными отметками, такими как среднее значение, нулевой уровень и перцентили (25-й, 50-й и 75-й), позволяет дополнительно анализировать поведение данных в отношении их статистических характеристик. Например, пересечения среднего значения могут указывать на моменты, когда тренд данных меняет направление, а пересечения с перцентилями — на моменты значительных изменений в поведении временного ряда.

Вместе эти параметры обеспечивают комплексную картину, позволяя анализировать временные ряды не только с точки зрения общей статистики, но и с учетом динамических особенностей их изменения. Это, в свою очередь, позволяет выделить группы схожих временных рядов, которые могут быть обработаны отдельными моделями для повышения точности прогнозирования.

Таким образом, каждый временной ряд будет представлен в виде 15-мерного вектора, состоящего из следующих ключевых характеристик:

- x_1 – минимальное значение;
- x_2 – максимальное значение;
- x_3 – среднее значение;
- x_5 – количество пиков;
- x_6 – количество впадин;
- x_7 – дисперсия элементов;
- x_8 – площадь графика;

- x_9 – сумма квадратов производной;
- x_{10} – коэффициенты ARIMA;
- x_{11} – нулевой уровень;
- x_{12} – 25-й перцентиль;
- x_{13} – 50-й перцентиль;
- x_{14} – 75-й перцентиль;
- x_{15} – показатель тренда.

Этап 3

После выделения параметров данные подвергаются кластеризации. Для этого используется метод DBSCAN [13], при этом данные предварительно масштабируются с помощью нормализации [14] для того, чтобы параметры, имеющие различные масштабы, не влияли на результаты кластеризации. Кластеризация позволяет группировать временные ряды с схожими характеристиками, что создает основу для создания отдельных прогнозных моделей для каждой группы.

Этап 4

На следующем этапе для каждого кластера обучается отдельная модель на основе LSTM. Это позволяет учитывать специфику каждого кластера, улучшая точность прогнозов.

Использование модели

При прогнозировании для нового временного ряда сначала определяется его кластер с помощью модели DBSCAN, а затем для соответствующего кластера используется обученная LSTM-модель для предсказания значений временного ряда.

Таким образом, предложенная методология решает проблему кластеризации и прогнозирования временных рядов, учитывая индивидуальные особенности данных и позволяя более точно моделировать динамику для различных типов временных рядов

Вычислительные эксперименты

Для проведения вычислительных экспериментов использовались данные, полученные с сенсоров инсулиновых помп, фиксирующих текущий уровень глюкозы в крови (SGV), количество активного инсулина (IOB Sum) и потребление углеводов (COB). Эти данные имеют непосредственную связь с динамикой изменения уровня глюкозы, и их совместное использование позволяет создавать более информативные модели для прогнозирования уровня глюкозы, чем использование отдельных параметров.

Данные для анализа были извлечены из логов сенсоров инсулиновых помп, которые включают временные метки, значения измерений и другие параметры. Для обработки данных использовались алгоритмы, основанные на регулярных выражениях, которые позволили точно извлечь нужные показатели. Эти данные затем были структурированы в табличный формат для удобства дальнейшего анализа и машинного обучения, что минимизировало возможные ошибки на этапе подготовки данных. Фрагмент сформированного смешанного временного ряда, используемого в исследовании, приведен в Таблице 1.

Обработанные данные сначала делились на более мелкие временные ряды длительностью 12 часов. Для каждого такого временного ряда определялись ключевые параметры:

1. Минимальное значение;
2. Максимальное значение;
3. Среднее значение;
4. Медианное значение;
5. Количество пиков;
6. Количество впадин;
7. Дисперсия элементов;
8. Площадь под графиком;
9. Сумма квадратов производной;
10. Коэффициенты ARIMA;
11. Нулевой уровень;
12. 25-й перцентиль;
13. 50-й перцентиль;
14. 75-й перцентиль;
15. Показатель тренда.

Таблица 1 – Фрагмент сформированного временного ряда

Table 1 – Fragment of the generated time series

SGV	IOB Sum	COB
83,71	6,15	9,91
83,72	6,03	9,79
83,92	2,54	9,57
84,17	2,42	9,27
84,61	2,16	8,98
84,95	2,06	8,81
85,55	2,06	8,79
86,23	1,94	8,82
86,37	1,65	8,75

В результате обработки обучающей выборки данных, сформированной на основе измерений, выполненных в течение восьми дней, было выделено 16 временных рядов. Для каждого временного ряда были рассчитаны ключевые параметры. Пример рассчитанных параметров представлен в Таблице 2.

Данные параметров временных рядов затем подвергались кластеризации с использованием алгоритма DBSCAN, который позволяет выделять группы на основе плотности. Полученные кластеры отражали различные состояния пациента, такие как нормальный уровень глюкозы, гипогликемия или гипергликемия. Данные распределились по 4 кластерам. Пример распределения записей по кластерам представлен в Таблице 3.

В результате кластеризации временные ряды распределились по четырем кластерам. Кластер 0 включает ряды 5, 9. Кластер 1 состоит из рядов 11, 12, 13, 16. Кластер 2 включает ряды 1, 2, 3, 4, 7, 10, 14, 15. Кластер 3 включает ряд 8.

Таблица 2 – Параметры временных рядов

Table 2 – Time series parameters

Ряд	Min	Max	Среднее	Медиана	Дисперсия	Площадь	…	Тренд
1	76	134,8	93,4971	87	268,598	93471,5	…	-0,04052
2	97	134	115,051	111,1	128,165	114904,5	…	-0,02207
3	84	135	110,856	113	312,149	110720,5	…	-0,05264
4	84	132	109,243	110	269,784	109129,5	…	-0,02338
5	55	160	112,751	119,5	1190,361	112664	…	-0,0854
6	85	163	135,128	142	616,3681	134903	…	-0,05
7	88	160	126,172	127	270,434	126040	…	0,05591
8	160	220	186,266	180,1	343,444	186047,5	…	0,05917
9	67	187	112,154	111	1063,401	112211	…	-0,01012
10	102	156	125,127	127,1	214,733	125057,5	…	-0,0197
11	137	176	158,835	157,4	116,403	158638	…	0,01230
12	145	175	159,736	162	75,564	159593,5	…	0,02604
13	135	176	146,844	142	121,029	146726,5	…	-0,02375
14	100	137	115,315	113	164,926	115262	…	0,00556
15	105	156	125,608	127	217,192	125542	…	0,02208
16	133	157	143,348	143	57,191	77587	…	-0,03487

Таблица 3 – Параметры временных рядов, распределенные по кластерам

Table 3 – Time series parameters distributed by clusters

Ряд	Min	Max	Среднее	Медиана	Дисперсия	Площадь	…	Тренд	Кластер
1	76	134,8	93,4971	87	268,598	93471,5	…	-0,04052	2
2	97	134	115,051	111,1	128,165	114904,5	…	-0,02207	2
3	84	135	110,856	113	312,149	110720,5	…	-0,05264	2
4	84	132	109,243	110	269,784	109129,5	…	-0,02338	2
5	55	160	112,751	119,5	1190,361	112664	…	-0,0854	0
6	85	163	135,128	142	616,3681	134903	…	-0,05	0
7	88	160	126,172	127	270,434	126040	…	0,05591	2
8	160	220	186,266	180,1	343,444	186047,5	…	0,05917	3
9	67	187	112,154	111	1063,401	112211	…	-0,01012	0
10	102	156	125,127	127,1	214,733	125057,5	…	-0,0197	2
11	137	176	158,835	157,4	116,403	158638	…	0,01230	1
12	145	175	159,736	162	75,564	159593,5	…	0,02604	1
13	135	176	146,844	142	121,029	146726,5	…	-0,02375	1
14	100	137	115,315	113	164,926	115262	…	0,00556	2
15	105	156	125,608	127	217,192	125542	…	0,02208	2
16	133	157	143,348	143	57,191	77587	…	-0,03487	1

Для каждого кластера обучалась отдельная нейронная сеть. При этом кластерный код не передавался в модель в качестве входного параметра, а использовался для разделения данных и создания специализированных моделей для каждого кластера. Для обучения нейронной сети с долгосрочной краткосрочной памятью (LSTM), эффективно моделирующей временные зависимости, из каждого кластера случайным образом выбирался один временной ряд, данные которого использовались для обучения. Такой подход позволял повысить точность прогнозирования уровня глюкозы в крови (SGV) для различных состояний пациента. Входными данными для модели служили параметры SGV, IOB Sum и COB, что обеспечивало учет динамики изменения уровня глюкозы.

Для оценки точности прогнозов использована средняя абсолютная процентная ошибка, широко применяемая в машинном обучении (MAPE, Mean Absolute Percentage Error) [15], определяемая формулой:

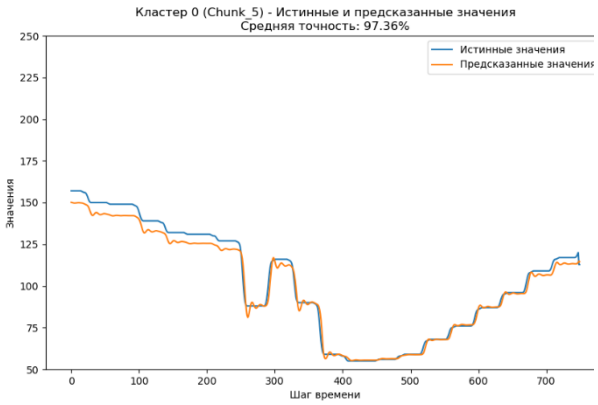
$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-\widehat{y_i}}{y_i}\right| * 100 \quad (4)$$

где y_i — фактическое значение на -м шаге, $\widehat{y_i}$ — предсказанное значение на i-м шаге, n — количество данных для анализа (количество временных шагов после обработки данных).

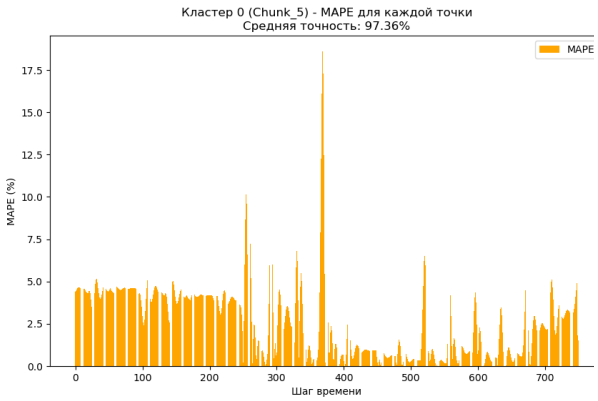
Точность определялась как величина, обратная ошибке:

$$\varepsilon = 100 - MAPE \qquad (5)$$

Для обучения LSTM-модели кластера 0 использовался **временной ряд 5**. Обучение показало среднюю точность 97,36%. Результаты обучения представлены на Рис. 1.



а



б

Рис. 1 – Результаты обучения для кластера 0: а) отклонение прогнозных значений от реальных; б) пошаговый уровень ошибки прогноза

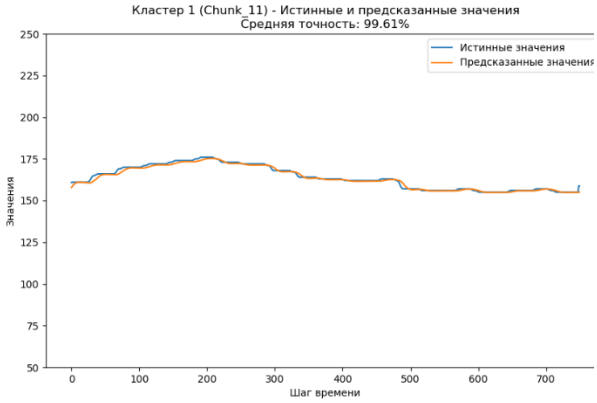
Fig. 1 – Training results for cluster 0: (a) Deviation of predicted values from real values; b) step-by-step level of forecast error

Для обучения LSTM-модели кластера 1 использовался временной ряд 11. Обучение показало среднюю точность 99,61%. Результаты обучения представлены на Рис. 2.

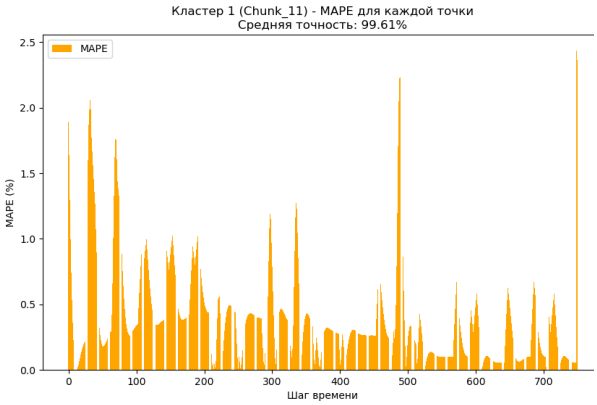
Для обучения LSTM-модели кластера 2 использовался временной ряд 1. Обучение показало среднюю точность 99,57%. Результаты обучения представлены на Рис. 3.

Для обучения LSTM-модели кластера 3 использовался временной ряд 8. Обучение показало среднюю точность 99,52%. Результаты обучения представлены на Рис. 4.

Процесс прогнозирования для новых данных включал несколько этапов. На первом этапе новый временной ряд, представляющий данные за сутки, был представлен в том же формате, что и обучающие данные.



а



б

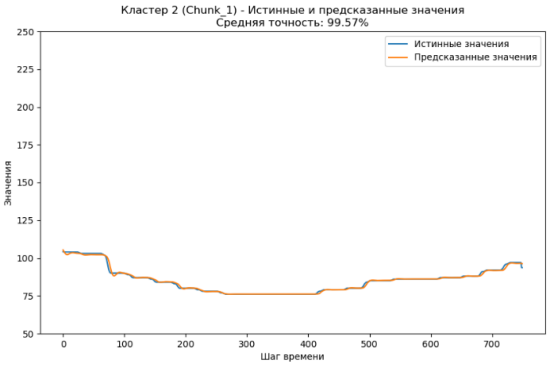
Рис. 2 – Результаты обучения для кластера 1: а) отклонение прогнозных значений от реальных; б) пошаговый уровень ошибки прогноза

Fig. 2 – Training results for cluster 1: (a) Deviation of predicted values from real values; b) step-by-step level of forecast error

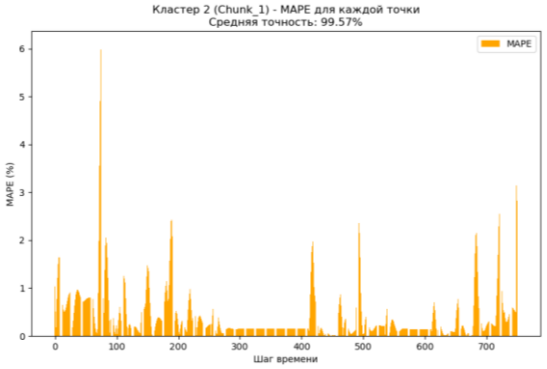
Далее для нового временного ряда определялся номер кластера с помощью обученной модели кластеризации DBSCAN, на основе тех же ключевых параметров, которые использовались при подготовке данных для кластеризации. После этого выбиралась соответствующая модель LSTM, обученная на данных данного кластера, и выполнялось прогнозирование уровня глюкозы (SGV) на заданный промежуток времени.

Выполним расчет для одного входного временного ряда. Для этого разделим тестовый временной ряд на 15 равных частей и выберем 3-й ряд. Этот 3-й временной ряд был отнесен к кластеру 3, и для его прогнозирования была использована LSTM-модель, обученная на данных, принадлежащих кластеру 3. Результат прогнозирования представлен на Рис. 5.

Обобщенные результаты прогнозирования представлены в Таблице 3.



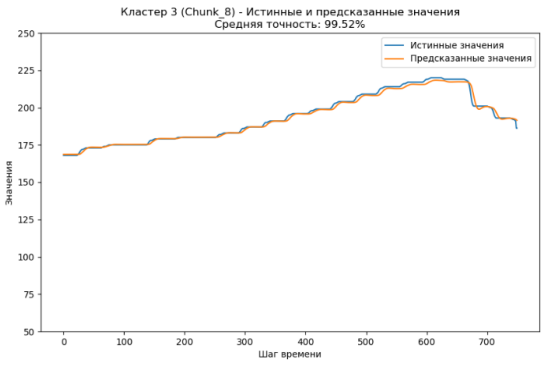
а



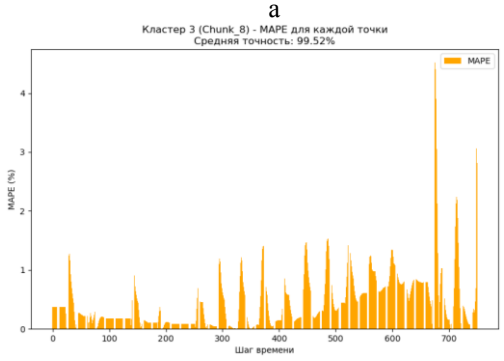
б

Рис. 3 – Результаты обучения для кластера 2: а) отклонение прогнозных значений от реальных; б) пошаговый уровень ошибки прогноза

Fig. 3 – Training results for cluster 2: (a) Deviation of predicted values from real values; b) step-by-step level of forecast error



а



б

Рис. 4 – Результаты обучения для кластера 3: а) отклонение прогнозных значений от реальных; б) пошаговый уровень ошибки прогноза

Fig. 4 - Training results for cluster 3: (a) Deviation of predicted values from real values; b) step-by-step level of forecast error

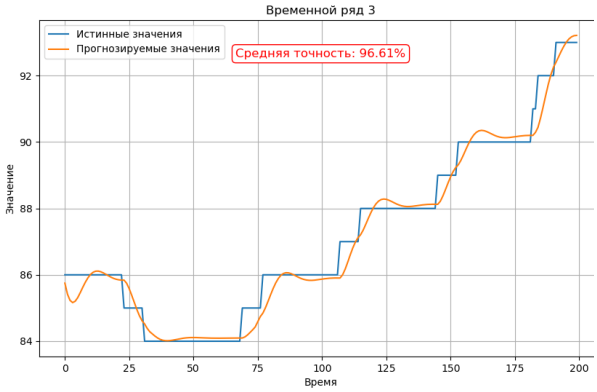


Рис. 5 – Результаты прогнозирования для нового временного ряда

Fig. 5 – Forecasting results for a new time series

Таблица 3 – Результаты прогнозирования

Table 3 – Forecasting results

Новый входной ряд	Кластер, к которому был определен ряд	Средняя точность прогноза
1	1	95,95533635%
2	2	95,63109678%
3	3	96,61094806%
4	2	95,81722369%
5	0	91,02030059%
6	0	92,40097836%
7	2	97,33380423%
8	2	91,63187927%
9	0	92,682793%
10	2	96,87877342%
11	1	99,30361861%
12	1	96,95888013%
13	2	96,73396784%
14	2	96,07275373%
15	3	95,8094856%

Сравнение эффективности с существующими методами

Для подтверждения преимущества разработанного метода в рамках данного исследования было проведено сравнение эффективности прогнозирования с другими методами, применимыми для кластеризации временных рядов. Сравнение предложенного метода осуществлялось по показателю точности, вычисляемому согласно формуле (5), с использованием следующих подходов:

- Обучение LSTM без предварительной кластеризации;

- Обучение LSTM с кластеризацией временных рядов на основе евклидова расстояния;
- Обучение LSTM с кластеризацией, основанной на вычислении параметров модели ARIMA.

Результаты тестирования трех методов, а также предложенного комбинированного метода представлены в Таблице 4.

Таблица 4 – Сравнение средней точки существующих методов

Table 4 – Comparison of the midpoint of existing methods

Модель	Средняя точность
LSTM без предварительной кластеризации	81,72%
LSTM с кластеризацией временных рядов на основе евклидова расстояния	83,46%
LSTM с кластеризацией, основанной на вычислении параметров модели ARIMA	91,69%
LSTM с кластеризацией на основе комбинированного подхода	95,23%

Рис. 6 демонстрирует графическое сравнение эффективности рассматриваемых методов.

Рис. 6 – Сравнение эффективности прогнозирования

Fig. 6 – Comparison of forecasting performance

Обсуждение

В этом исследовании основное внимание было уделено методам представления временных рядов через параметры для дальнейшего их использования в прогнозировании. Ключевым аспектом стало использование кластеризации для выявления схожих паттернов во временных рядах, что позволило создать более точные модели прогнозирования.

Результаты показали, что представление временных рядов через параметры с использованием кластеризации помогает значительно повысить точность прогноза. В частности, модели, обученные на параметризованных данных, показали снижение ошибки прогноза на 14% по сравнению с прогнозами, сделанными на исходных временных рядах без преобразования. Это подтверждает гипотезу о том, что представление данных через параметры позволяет скрыть избыточность и фокусироваться на ключевых характеристиках, что способствует лучшему прогнозированию.

При применении алгоритма DBSCAN для кластеризации временных рядов, результатом стала возможность выделить группы временных рядов с общими параметрическими характеристиками. Модели LSTM, обученные на этих кластерах, показали значительное улучшение точности прогноза, с уменьшением средней ошибки на 10%. Это говорит о том, что работа с параметризованными данными позволяет нейронным сетям более эффективно выявлять закономерности и делать более точные прогнозы.

Заключение

В данной работе предложена методология кластеризации временных рядов и адаптивного прогнозирования с использованием специализированных моделей на основе LSTM. Разработанный подход позволяет решать ключевые проблемы анализа временных рядов, такие как высокая вариативность данных, наличие нелинейных зависимостей и необходимость персонализации прогнозов.

Предложенный двухэтапный подход включает выделение параметров временных рядов с последующей кластеризацией методом DBSCAN и построение специализированных моделей для каждого кластера. Такой метод позволяет учитывать индивидуальные особенности временных рядов, улучшать точность прогнозов и эффективно справляться с гетерогенными и шумными данными.

Результаты вычислительных экспериментов, выполненных на данных инсулиновых помп, демонстрируют эффективность предложенного подхода. Кластеризация временных рядов позволила выделить группы с различными состояниями пациента, а адаптивные модели, обученные на данных этих групп, обеспечили более точное прогнозирование уровня глюкозы в крови.

Перспективы дальнейших исследований включают расширение набора параметров для кластеризации, использование методов самообучения для улучшения качества моделей, а также применение разработанного подхода к другим типам временных рядов в различных областях, таких как энергетика, медицина и финансы. Это позволит адаптировать предложенную методологию к задачам с более сложными структурами данных и увеличить её универсальность.

Предложенная методология демонстрирует потенциал для применения в реальных задачах анализа временных рядов, предоставляя гибкий и эффективный инструмент для работы с большими и сложными данными.

Литература

1.Ryabko, Daniil. (2020). Universal Time-Series Forecasting with Mixture Predictors. 10.1007/978-3-030-54304-4.

2. Puchkin, Nikita & Timofeev, Aleksandr & Spokoiny, Vladimir. (2020). Manifold-based time series forecasting. 10.48550/arXiv.2012.08244.
3. Kalpakis, K. & Gada, Dhiral & Puttagunta, Vasundhara. (2002). Distance Measures for Effective Clustering of ARIMA Time-Series.
4. Berthold, Michael & Höppner, Frank. (2016). On Clustering Time Series Using Euclidean Distance and Pearson Correlation. 10.48550/arXiv.1601.02213.
5. W. Wang, G. Lyu, Y. Shi and X. Liang, "Time Series Clustering Based on Dynamic Time Warping," *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, Beijing, China, 2018, pp. 487-490, doi: 10.1109/ICSESS.2018.8663857.
6. Holan, Scott & Ravishanker, Nalini. (2018). Time series clustering and classification via frequency domain methods. Wiley Interdisciplinary Reviews: Computational Statistics. 10. e1444. 10.1002/wics.1444.
7. L. Xingye and T. Tian, "Time series recognition based on wavelet transform and Fourier transform," *2010 IEEE Symposium on Industrial Electronics and Applications (ISIEA)*, Penang, Malaysia, 2010, pp. 722-726, doi: 10.1109/ISIEA.2010.5679372.
8. Kalpakis, K. & Gada, Dhiral & Puttagunta, Vasundhara. (2001). Distance Measures for Effective Clustering of ARIMA Time-Series. Proceedings - IEEE International Conference on Data Mining, ICDM. 273-280. 10.1109/ICDM.2001.989529.
9. Tavakoli, Neda & Siami Namini, Sima & Khanghah, Mahdi & Soltani, Fahimeh & Siami Namin, Akbar. (2020). Clustering Time Series Data through Autoencoder-based Deep Learning Models. 10.48550/arXiv.2004.07296.
10. Trosten, Daniel & Strauman, Andreas & Kampffmeyer, Michael & Jenssen, Robert. (2018). Recurrent Deep Divergence-based Clustering for simultaneous feature learning and clustering of variable length time series. 10.48550/arXiv.1811.12050.
11. Huang, Fanling & Deng, Yangdong. (2023). TCGAN: Convolutional Generative Adversarial Network for Time Series Classification and Clustering. 10.48550/arXiv.2309.04732.
12. Tawakuli, Amal & Havers, Bastian & Gulisano, Vincenzo & Kaiser, Daniel. (2024). Survey: Time-Series Data Preprocessing: A Survey and an Empirical Analysis. Journal of Engineering Research. 10.1016/j.jer.2024.02.018.
13. C. -F. Chien and B. Suwattananuruk, "Density-Based Spatial Clustering of Applications With Noise (DBSCAN) for Probe Card Production for Advanced Quality Control of Wafer Probing Test," in IEEE Transactions on Semiconductor Manufacturing, vol. 37, no. 4, pp. 567-575, Nov. 2024, doi: 10.1109/TSM.2024.3468000.
14. Gamakumara, Puwasala & Santos-Fernández, Edgar & Talagala, Priyanga & Hyndman, Rob & Mengersen, Kerrie & Leigh, Catherine. (2023). Conditional normalization in time series analysis. 10.48550/arXiv.2305.12651.
15. Statology. (2020). Mean Absolute Percentage Error (MAPE) in Machine Learning. Retrieved from https://www.statology.org/mape-in-machine-learning/

References

1. Ryabko, Daniil. (2020). Universal Time-Series Forecasting with Mixture Predictors. 10.1007/978-3-030-54304-4.

2. Puchkin, Nikita & Timofeev, Aleksandr & Spokoiny, Vladimir. (2020). Manifold-based time series forecasting. 10.48550/arXiv.2012.08244.
3. Kalpakis, K. & Gada, Dhiral & Puttagunta, Vasundhara. (2002). Distance Measures for Effective Clustering of ARIMA Time-Series.
4. Berthold, Michael & Höppner, Frank. (2016). On Clustering Time Series Using Euclidean Distance and Pearson Correlation. 10.48550/arXiv.1601.02213.
5. W. Wang, G. Lyu, Y. Shi and X. Liang, "Time Series Clustering Based on Dynamic Time Warping," *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, Beijing, China, 2018, pp. 487-490, doi: 10.1109/ICSESS.2018.8663857.
6. Holan, Scott & Ravishanker, Nalini. (2018). Time series clustering and classification via frequency domain methods. Wiley Interdisciplinary Reviews: Computational Statistics. 10. e1444. 10.1002/wics.1444.
7. L. Xingye and T. Tian, "Time series recognition based on wavelet transform and Fourier transform," *2010 IEEE Symposium on Industrial Electronics and Applications (ISIEA)*, Penang, Malaysia, 2010, pp. 722-726, doi: 10.1109/ISIEA.2010.5679372.
8. Kalpakis, K. & Gada, Dhiral & Puttagunta, Vasundhara. (2001). Distance Measures for Effective Clustering of ARIMA Time-Series. Proceedings - IEEE International Conference on Data Mining, ICDM. 273-280. 10.1109/ICDM.2001.989529.
9. Tavakoli, Neda & Siami Namini, Sima & Khanghah, Mahdi & Soltani, Fahimeh & Siami Namin, Akbar. (2020). Clustering Time Series Data through Autoencoder-based Deep Learning Models. 10.48550/arXiv.2004.07296.
10. Trosten, Daniel & Strauman, Andreas & Kampffmeyer, Michael & Jenssen, Robert. (2018). Recurrent Deep Divergence-based Clustering for simultaneous feature learning and clustering of variable length time series. 10.48550/arXiv.1811.12050.
11. Huang, Fanling & Deng, Yangdong. (2023). TCGAN: Convolutional Generative Adversarial Network for Time Series Classification and Clustering. 10.48550/arXiv.2309.04732.
12. Tawakuli, Amal & Havers, Bastian & Gulisano, Vincenzo & Kaiser, Daniel. (2024). Survey: Time-Series Data Preprocessing: A Survey and an Empirical Analysis. Journal of Engineering Research. 10.1016/j.jer.2024.02.018.
13. C. -F. Chien and B. Suwattananuruk, "Density-Based Spatial Clustering of Applications With Noise (DBSCAN) for Probe Card Production for Advanced Quality Control of Wafer Probing Test," in IEEE Transactions on Semiconductor Manufacturing, vol. 37, no. 4, pp. 567-575, Nov. 2024, doi: 10.1109/TSM.2024.3468000.
14. Gamakumara, Puwasala & Santos-Fernández, Edgar & Talagala, Priyanga & Hyndman, Rob & Mengersen, Kerrie & Leigh, Catherine. (2023). Conditional normalization in time series analysis. 10.48550/arXiv.2305.12651.
15. Statology. (2020). Mean Absolute Percentage Error (MAPE) in Machine Learning. Retrieved from https://www.statology.org/mape-in-machine-learning/

© **3. 3. Мингалиев –** аспирант, Казанский национальный исследовательский технический университет им. А.Н. Туполева, Казань, Россия, zaid97abyi@gmail.com

© **Zaid Mingaliev –** PhD-student of Kazan National Research Technical University named after A.N. Tupolev, Kazan, Russia, zaid97abyi@gmail.com