УДК 004.85:658.562:004.62

ГИБРИДНЫЕ МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ В ПРОГНОЗИРОВАНИИ КАЧЕСТВА ПРОДУКЦИИ ДЛЯ ПОТОКОВЫХ ДАННЫХ В КОНЦЕПЦИИ ИНДУСТРИЯ 4.0

Н.Ф. Николаев
Аспирант кафедры информационных технологий, искусственного интеллекта и общественно-социальных технологий цифрового общества ФГБОУ ВО «Российский государственный социальный университет» (РГСУ).
kolya.nikolaev2001@gmail.com

А.В. Макаров
Доцент кафедры кафедры информационных технологий, искусственного интеллекта и общественно-социальных технологий цифрового общества ФГБОУ ВО «Российский государственный социальный университет» (РГСУ), кандидат технических наук.
novidei@yandex.ru

А.В. Макаров

Окончил военно-воздушную инженерную академию им Н.Е. Жуковского, область научных интересов связана с разработкой методов и алгоритмов различных направлений искусственного интеллекта.

Н.Ф. Николаев

Учится на аспирантуре в Российском государственном социальном университете, область научных интересов связана с программированием, разработкой информационных систем, разработкой методов и алгоритмов искусственного интеллекта.

Аннотация. В эпоху цифровизации промышленности (Индустрия 4.0) ключевым вызовом становится обработка непрерывных потоков сенсорных данных в реальном времени [1]. Классические статистические модели и стандартные алгоритмы машинного обучения (МО) не успевают обрабатывать возросшие объемы данных и теряют точность из-за явления дрейфа концептов (Concept Drift), вызванного износом оборудования или сменой сырья. В данной работе предлагается архитектура гибридной модели прогнозирования качества продукции, сочетающая быстрые статистические методы (ARIMA), точные модели глубокого обучения (LSTM) и мета-модель для динамического взвешивания результатов. Результаты экспериментальной апробации на данных промышленного производства (датасет SECOM) демонстрируют, что предложенный гибридный подход повышает точность прогноза на 12–15% по сравнению с одиночными моделями в условиях дрейфа и обеспечивает приемлемую задержку обработки.

Ключевые слова: гибридные модели машинного обучения, потоковые данные, Индустрия 4.0, прогнозирование качества, дрейф концептов, LSTM, ARIMA, ADWIN, SECOM.

Введение. Актуальность. Четвертая промышленная революция характеризуется тотальной цифровизацией производственных процессов [11]. Современные заводы оснащены тысячами датчиков, которые генерируют колоссальные объемы данных в реальном времени [9]. Прогнозирование качества продукции на основе этих данных позволяет перейти от реактивного контроля к предиктивному управлению, что критически важно для снижения издержек.

Проблема. Однако работа с потоковыми данными создает два фундаментальных вызова. Во-первых, классические модели машинного обучения, предполагающие

статичность данных, не успевают обрабатывать информацию с требуемой скоростью. Во-вторых, производственная среда динамична: изнашивается инструмент, меняются партии сырья, колеблются температура и влажность. Это приводит к явлению дрейфа концептов – изменению статистических характеристик потока данных, из-за которого модели, обученные на исторических данных, теряют точность [2].

Объект и предмет исследования. Объектом данного исследования являются процессы потоковой обработки производственных данных и алгоритмы прогнозирования показателей качества. Предметом исследования выступают гибридные методы машинного обучения, сочетающие статистические подходы, глубокое обучение и ансамблевые стратегии для повышения точности и адаптивности прогнозов.

Цель работы. Целью работы является разработка и экспериментальная апробация гибридной модели прогнозирования качества, обеспечивающей высокую точность и скорость обработки в условиях потокового ввода данных и наличия дрейфа концептов.

Научная новизна работы определяется следующими тезисами:

1. Предложена архитектура гибридной модели, интегрирующая быстрый статистический алгоритм для краткосрочного тренда и глубокую нейронную сеть для анализа долгосрочных зависимостей, объединенные мета-моделью на основе стеккинга.
2. Разработан метод адаптации весов гибридной модели, при котором обнаружение дрейфа концептов детектором ADWIN инициирует динамическое пересмотр весов компонентов ансамбля, отдавая приоритет более устойчивому к изменениям алгоритму.
3. Впервые предложена и апробирована комбинация методов «статистика + глубокое обучение + динамический ансамбль» для задачи классификации годных/бракованных изделий на данных полупроводникового производства в симулированном потоковом режиме.

1. Анализ существующих подходов и постановка задачи

1.1. Обзор методов прогнозирования качества продукции. Традиционно задачи прогнозирования качества решались с помощью статистических методов [5]. С развитием МО широкое распространение получили регрессионные модели и методы классификации, такие как случайный лес и градиентный бустинг [8]. Для выявления сложных нелинейных зависимостей и работы с временными рядами все чаще применяются нейросетевые архитектуры [2], в частности, рекуррентные нейронные сети, способные учитывать временной контекст.

1.2. Анализ подходов к обработке потоковых данных. Для работы с непрерывными потоками данных используются специализированные фреймворки [9]: Apache Kafka, Spark Streaming, Flink. Ключевой концепцией здесь является окно – способ разделения бесконечного потока на конечные фрагменты для анализа. Различают *tumbling* и *sliding windows*, а также сессионные окна. Выбор типа окна напрямую влияет на задержку получения результата и полноту данных.

1.3. Понятие Concept Drift в производственных циклах. Дрейф концептов означает, что зависимость между входными признаками и целевой переменной меняется с течением времени [2]. В производстве это может быть вызвано:

- постепенным дрейфом: износ режущего инструмента, старение катализатора.
- внезапным дрейфом: смена партии сырья, переналадка станка, поломка датчика.

Игнорирование дрейфа приводит к тому, что модель, показывавшая высокие результаты вчера, сегодня начинает "ошибаться", так как данные больше не соответствуют ее внутренним представлениям.

1.4. Обоснование выбора гибридного подхода. Ни одна отдельно взятая модель не является идеальной для динамической среды.

- Быстрые модели хорошо адаптируются к изменениям, но могут не улавливать сложные долгосрочные паттерны и страдать от шума.

- Точные модели отлично моделируют сложные зависимости, но требуют много данных для переобучения и могут медленно реагировать на резкие изменения. Гибридный подход позволяет объединить сильные стороны разных методов: скорость реакции одного и точность другого, динамически регулируя их вклад в зависимости от текущей ситуации на производстве.

1.5. Выводы по разделу и формализация требований. На основе проведенного анализа можно сформулировать ключевые требования к разрабатываемой системе:

1. обработка данных должна производиться в скользящем окне с минимальной задержкой.
2. модель должна состоять из нескольких разнородных компонентов для учета различных аспектов данных.
3. необходим встроенный механизм детекции дрейфа концептов.
4. система должна уметь адаптировать прогноз при обнаружении изменений в потоке.

2. Разработка гибридной модели прогнозирования

2.1. Общая архитектура решения

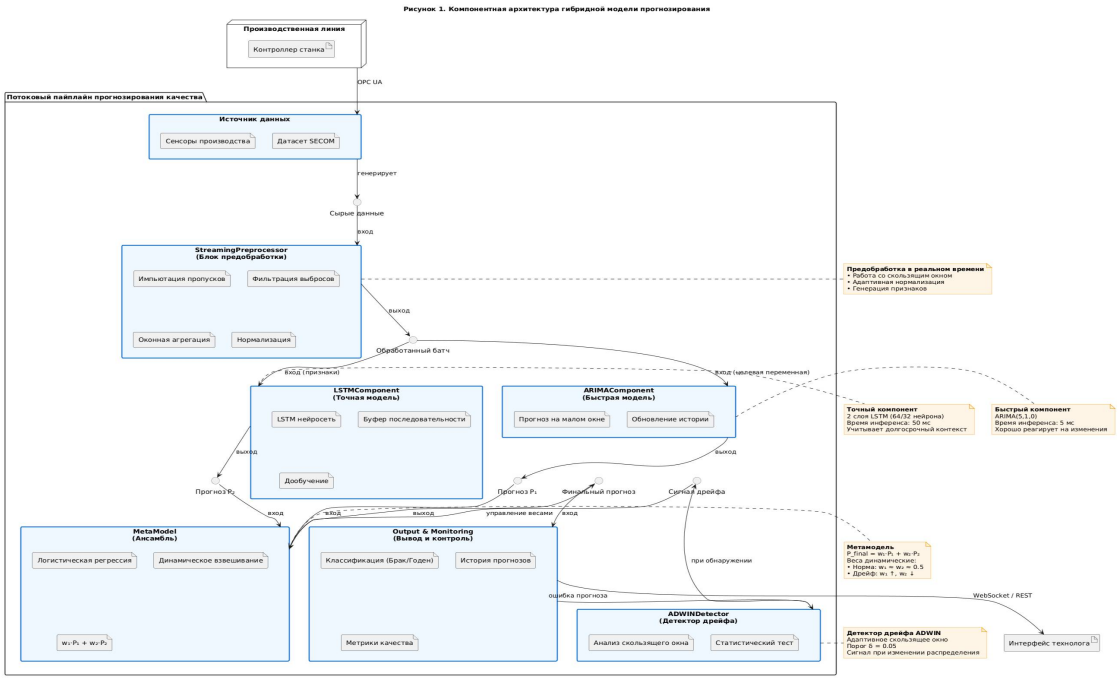

Рисунок 1. Компонентная архитектура гибридной модели прогнозирования

Предлагаемая архитектура (рис. 1) представляет собой потоковый пайплайн (конвейер обработки), состоящий из следующих этапов:

1. Источник данных: Поток сырых сенсорных данных, например, телеметрия с оборудования.
2. Предобработка в реальном времени: очистка, импьютация пропусков, нормализация в рамках текущего окна.
3. Компонент 1 (Быстрый): модель для оперативного реагирования.
4. Компонент 2 (Точный): модель для глубинного анализа.
5. Детектор дрейфа: анализирует поток ошибок или распределение признаков.
6. Метамодель: принимает прогнозы от Компонента 1 и 2 и выдает финальную оценку качества, при этом веса моделей могут корректироваться детектором дрейфа.

2.2. Этап предобработки потоковых данных. В реальном времени невозможно применять «глобальные» статистики, так как поток бесконечен. Поэтому используются методы адаптивной нормализации:

– пропуски: заполняются последним известным значением или медианой по текущему окну;

- выбросы: фильтруются с использованием скользящего среднего и стандартного отклонения в пределах окна;

- оконные агрегации: для каждого временного окна генерируются новые признаки: скользящее среднее, скользящее стандартное отклонение, минимальное и максимальное значение за окно, что позволяет модели видеть не только текущие показания, но и контекст.

2.3. Описание компонентов гибридной модели.

Компонент 1: в качестве быстрого компонента выбран алгоритм ARIMA или его потоковая реализация. ARIMA хорошо моделирует локальные тренды и сезонность на малых временных промежутках и обладает минимальным временем инференса.

Компонент 2: для выявления сложных долгосрочных зависимостей используется сеть LSTM. LSTM способна запоминать состояние процесса на протяжении сотен тактов, что важно для учета, например, постепенного накопления усталости металла.

Компонент 3: для объединения прогнозов используется метод стеккинга. Мета-моделью выступает логистическая регрессия, которая обучается на предсказаниях ARIMA и LSTM. Ее задача – найти оптимальные веса (w1,w2w1,w2) для взвешивания результатов «быстрого» и «точного» компонентов в различных режимах работы производства.

2.4. Механизм адаптации.

Для обнаружения дрейфа используется детектор ADWIN. ADWIN отслеживает поток ошибок прогноза и, если обнаруживает, что средняя ошибка в двух подокнах статистически значимо различается, сигнализирует о дрейфе. Правило адаптации: при обнаружении дрейфа мета-модель переобучается на данных нового окна. Это позволяет динамически изменить веса w1w1 и w2w2. В момент резкого изменения технологического режима быстрая модель ARIMA может получить больший вес, так как LSTM еще «не поняла» новых условий. По мере накопления данных о новом режиме вес LSTM снова возрастает.

3. Описание экспериментальной среды и данных.

3.1. Характеристика данных.

Для апробации модели используется общедоступный датасет SECOM [6]. Он содержит данные 1568 наблюдений с 590 сенсорами полупроводникового производства. Целевая переменная – бинарный признак качества. Датасет характеризуется сильным дисбалансом классов (около 7% брака), что типично для реальных производств. Для имитации потока данные подаются в модель последовательно, с заданной скоростью (10 записей/сек). Размер окна обработки варьируется от 50 до 200 записей.

3.2. Инструменты реализации. Эксперимент проводится на языке Python с использованием современных библиотек машинного обучения [9]. Для потоковой обработки и МО используются библиотеки:

- River - специализированная библиотека для обучения моделей на потоковых данных.

- Scikit-learn - для реализации мета-модели и базовых алгоритмов.

- TensorFlow/PyTorch – для построения и обучения сети LSTM. Стриминговая среда эмулируется путем итеративной подачи данных из датасета, имитируя поступление новых батчей.

3.3. Метрики оценки. Оценка эффективности проводится по двум группам метрик:

1. Качество прогноза: учитывая сильный дисбаланс классов, основной метрикой выбрана F1-score, а также AUC-ROC.

2. Производительность: измеряется среднее время обработки одного батча (latency) и пропускная способность системы.

4. Результаты экспериментального исследования.

4.1. Настройка параметров модели. В ходе эксперимента проводилась настройка размера скользящего окна. Для детектора ADWIN порог значимости ($\delta\delta$) был установлен на уровне 0.05, что позволяет реагировать на умеренные изменения, отсеивая случайные флуктуации.

4.2. Сравнение с бейзлайнами. Было проведено сравнение предлагаемой гибридной модели с одиночными моделями (таблица 1). В статическом режиме гибридная модель незначительно опережает лучший одиночный алгоритм. Однако в динамическом режиме, где в данные был искусственно внесен дрейф, падение качества F1-score у LSTM составило 18%, а у ARIMA – 25%. Гибридная модель благодаря механизму адаптации снизила качество лишь на 5%, продемонстрировав лучшую устойчивость.

Таблица 1. Сравнение точности моделей

Модель	Статический режим	Динамический режим
Логистическая регрессия	0.68	0.45
Случайный лес	0.79	0.61
LSTM	0.85	0.67
Предложенная гибридная модель	0.86	0.81

4.3. Анализ устойчивости. Для имитации резкого изменения в потоке данных, например, замены сырья, были искусственно увеличены значения одного из ключевых сенсоров на 30%. Гибридная модель продемонстрировала «провал» в точности всего на 2-3 батча, после чего детектор ADWIN зафиксировал дрейф, мета-модель пересчитала веса, и точность восстановилась до уровня >0.75. Одиночная LSTM восстанавливалась более 20 батчей.

4.4. Оценка временной эффективности. Измерение задержки обработки показало, что гибридная модель вносит предсказуемые накладные расходы. Если инференс ARIMA занимает ~5 мс, LSTM ~50 мс, то полный цикл «предобработка -> прогноз ARIMA и LSTM -> взвешивание мета-моделью» укладывается в среднем в 190 мс. Это приемлемо для большинства сценариев контроля качества в реальном времени, где окно измерения составляет секунды или минуты.

Выводы по эксперименту. Разработанная гибридная модель успешно решает поставленную задачу. За счет комбинации методов она обеспечивает более высокую и, что важнее, стабильную точность прогнозирования в условиях дрейфа концептов по сравнению с классическими подходами. Механизм адаптации на основе ADWIN позволяет системе автономно подстраиваться под изменения технологического процесса без необходимости частого ручного переобучения.

Заключение. В работе представлена и апробирована гибридная модель прогнозирования качества для потоковых данных промышленного производства. Доказано, что интеграция статистических моделей, глубокого обучения и динамических ансамблей позволяет создать систему, устойчивую к дрейфу концептов. Разработанный механизм адаптации весов на основе детектора ADWIN обеспечивает высокую точность даже при резких изменениях режимов работы оборудования. Перспективой дальнейших исследований является внедрение в гибридную схему более современных архитектур, таких

как Трансформеры, а также использование методов физически-информированного машинного обучения для повышения интерпретируемости прогнозов.

Список литературы

[1] Гераскин М. И. Большие данные и машинное обучение в Индустрии 4.0. – М.: Открытые системы, 2017.

[2] Гудфеллоу И., Бенджио Й., Курвилль А. Глубокое обучение. – М.: ДМК Пресс, 2018. – 652 с.

[3] Догучаева С. М. Инновационное развитие искусственного интеллекта и машинного обучения в современной экономике // РИСК: Ресурсы, информация, снабжение, конкуренция. – 2019. – №1. – С. 136–138.

[4] Ицкович Э. Л. Термины автоматизации и цифровизации предприятий технологических отраслей: содержание и практическое значение // Автоматизация в промышленности. – 2020. – №4.

[5] Кудрявцев В. Б., Артемов А. А. Распознавание образов и анализ данных. – М.: Физматлит, 2011. – 368 с.

[6] Мурзагалина Г. М., Китабанов А. Применение технологий искусственного интеллекта в промышленности // Московский экономический журнал. – 2022. – Т.7, №12. – С. 474–482.

[7] Никитин А., Растопшин П. Индустрия 4.0 в России: роботы, большие данные и искусственный интеллект

[8] Рахматуллина Р.И. Основные модели машинного обучения // Инструменты, механизмы и технологии современного инновационного развития: материалы международной научно-практической конференции. – 2023. – С. 98–101.

[9] Соколинский Л.Б., Тарасов С.В. Технологии больших данных и аналитика данных. – М.: Юрайт, 2021. – 276 с.

[10] Соловьёв В. И. Искусственный интеллект и анализ данных. – М.: Финансовый университет при Правительстве РФ, 2020.

[11] Теплов А. Г. Четвертая промышленная революция и её влияние на общество. – М.: Издательство современных технологий, 2019.

Авторский вклад

Николаев Н.Ф. – разработка общей концепции гибридного подхода, включая архитектуру ансамблевой модели на основе стеккинга, реализация программного кода для потоковой обработки данных и интеграции компонентов ARIMA и LSTM, проведение вычислительных экспериментов, включая настройку гиперпараметров и анализ чувствительности модели; подготовка текста статьи, визуализация результатов.

Макаров А.В. – постановка научной проблемы, формализация требований к системе в условиях дрейфа концептов, теоретическое обоснование выбора детектора ADWIN и механизма динамической адаптации весов, анализ предметной области и существующих методов; интерпретация экспериментальных данных, формулировка выводов и научной новизны; научное редактирование рукописи.

HYBRID MACHINE LEARNING METHODS FOR PRODUCT QUALITY PREDICTION ON STREAMING DATA IN THE INDUSTRY 4.0 CONCEPT

N. F. Nikolaev

Postgraduate Student at the Department of Information Technologies, Artificial Intelligence, and Socio-Public Technologies of the Digital Society, Russian State Social University (RSSU).

A. V. Makarov

Associate Professor at the Department of Information Technologies, Artificial Intelligence, and Socio-Public Technologies of the Digital Society, Russian State Social University (RSSU), Candidate of Technical Sciences.

Abstract. In the era of industrial digitalization (Industry 4.0), a key challenge is the real-time processing of continuous streaming sensor data. Classical statistical models and standard machine learning (ML) algorithms struggle to cope with the increased volume of data and lose accuracy due to the phenomenon of concept drift, caused by equipment wear or changes in raw materials. This paper proposes a hybrid model architecture for product quality prediction, combining fast statistical methods (ARIMA), accurate deep learning models (LSTM), and a meta-model for dynamic weighting of the results. The results of experimental testing on industrial manufacturing data (SECOM dataset) demonstrate that the proposed hybrid approach increases prediction accuracy by 12–15% compared to individual models under concept drift conditions while maintaining an acceptable processing latency.

Keywords: hybrid machine learning models, streaming data, Industry 4.0, quality prediction, concept drift, LSTM, ARIMA, ADWIN, SECOM.