Методика предварительного отбора различных последовательностей данных на основе относительного отклонения для формирования обучающих выборок в задачах машинного обучения

Е.А. Дудалова, О.В. Соловьева, С.А. Соловьев

Казанский государственный энергетический университет, Казань

Аннотация: В настоящем исследовании представлена методика предварительной обработки последовательностей данных, направленная на выявление и группировку различных файлов данных для последующего использования при обучении нейронных сетей. Предложен алгоритм сравнения файлов на основе относительного отклонения значений признаков с учётом граничных случаев (нулевые и близкие к нулю значения). Реализация включает параллельную обработку для повышения производительности и генерацию детализированных отчётов. Метод протестирован на наборе данных, содержащем 10000 файлов с показателями химического процесса в лабораторном реакторе. Результаты показывают эффективность метода в выявлении стационарных участков и формировании сбалансированных обучающих выборок.
Ключевые слова: предобработка данных, относительное отклонение, машинное обучение, параллельные вычисления, группировка файлов, математическое моделирование вычислительная гидродинамика.

Введение

Технологические процессы нефтехимии составляют фундаментальный базис современной глобальной экономики. Ввиду сложности технологий для ускорения исследования процессов используют методы численного моделирования, которые позволяют решать сложные математические уравнения, описывающие физические и химические процессы, с помощью высокопроизводительной вычислительной техники. В нефтехимии, где процессы часто протекают при экстремальных температурах и давлениях, а эксперименты дороги и опасны, такой подход стал основным инструментом [1]. Применения численного моделирования охватывает всю цепочку – от проектирования оборудования до оптимизации производства и обеспечения безопасности.

Машинное обучение [2 – 4] в численном моделировании не отменяет классические методы, а дополняет и усиливает их, позволяя решать задачи, которые были слишком сложны, дороги или медленны для традиционных

подходов. Полноценное численное моделирование нефтехимического реактора или процесса может занимать часы, дни или даже недели на мощных кластерах. Это делает невозможным, например, оперативную оптимизацию или анализ. На основе ограниченного количества заранее проведенных расчетов можно обучить быструю суррогатную модель (например, нейронную сеть). После обучения суррогатная модель выдает результат за миллисекунды, позволяя проводить оптимизацию в реальном времени параметров.

В задачах машинного обучения качество обучающей выборки напрямую влияет на обобщающую способность модели [5]. Это особенно актуально при работе с большими наборами данных, сгенерированными в результате численного моделирования или физических экспериментов. Такие наборы часто содержат значительную избыточность [6]: последовательные временные срезы состояния системы могут быть практически идентичными, особенно в стационарных или квазистационарных режимах. Использование всех таких данных при обучении приводит к неоптимальному расходу вычислительных ресурсов, увеличению времени обучения и риску переобучения на дублирующих примерах.

Цель настоящей работы – разработка и реализация высокопроизводительного алгоритма для автоматической идентификации и группировки практически идентичных файлов в крупномасштабном наборе данных, полученном в результате моделирования химического процесса гидрирования ацетилена в лабораторном реакторе [1]. Алгоритм основан на попарном сравнении соседних во времени файлов с использованием метрики относительного отклонения по ключевым признакам. Для обработки малых значений реализован устойчивый метод с логарифмированием. Используется метод распараллеливания вычислений для максимизации производительности.

Исходные данные для машинного обучения

В качестве примера источника данных рассмотрим процесс очистки этилена от ацетилена [1]. Реактор представлял собой полую стальную трубку с внутренним диаметром 20 мм, а слой катализатора имеет длину 55 мм. Диапазон температур газа составляет 30–60°C при давлении 1 атм. Газовая фаза на входе в реактор состоит из компонент: аргон (24%), этилен (74,93%), водород (0,07%) и ацетилен (1%). Проходя через катализатор, ацетилен гидрируется до этилена. Помимо основной реакции, могут протекать и нежелательные побочные реакции, например, гидрирование этилена до этана.

$$C_2H_2 + H_2 \rightarrow C_2H_4 + 176\ \text{кДж/моль}, \quad (1)$$

$$C_2H_4 + H_2 \rightarrow C_2H_6 + 138\ \text{кДж/моль}. \quad (2)$$

Для проведения вычислительного эксперимента использован пакет программ ANSYS Fluent 19.2. Процесс нестационарный, временной шаг в расчетах численного моделирования выбран равным 0,005 с. На каждом временном шаге сохранялся файл с данными.

Исходные данные для машинного обучения представляют собой 10 000 текстовых файлов с именами вида data-XXXX. Здесь XXXX – номер итерации временного шага от 0001 до 10000). Каждый файл содержит шесть вещественных чисел, соответствующих температуре газа и нормированным значениям массового содержания компонентов газовой фазы в фиксированный момент времени для каждой итерации временного шага:

— temperature – температура;
— h2 – концентрация водорода;
— c2h6 – массовое содержание этана в газовой фазе;
— c2h4 – массовое содержание этилена в газовой фазе;
— c2h2 – массовое содержание ацетилена в газовой фазе;
— ar – массовое содержание аргона в газовой фазе.

Важной особенностью данных является то, что сумма концентраций всех компонентов (кроме температуры) строго равна единице. Это означает, что в процессе реакции часть компонентов может практически исчезать, принимая значения, близкие к машинному нулю ($<10^{-10}$). При вычислении относительных отклонений такие значения требуют особой обработки во избежание численной неустойчивости или ложных срабатываний. Формат данных: значения разделены пробелами, тип данных – float64. Общий объём данных – 10000 файлов, что даёт 9999 пар для сравнения соседних файлов.

Алгоритм отбора данных

Для решения задачи отбора данных предложен алгоритм, основанный на попарном сравнении соседних во времени файлов с использованием метрики относительного отклонения по каждому признаку. Выбор соседних файлов обусловлен постановкой будущей задачи численного моделирования с применением обучения нейронной сети: необходимо подать на вход файл с данными для временного шага t_n и получить на выходе предсказания параметров для временного шага t_{n+1}.

Для обеспечения устойчивости при работе с малыми значениями массового содержания компонентов газовой фазы реализован метод, использующий логарифмическое масштабирование. Для достижения высокой производительности при обработке десятков тысяч файлов применено многопроцессорное распараллеливание этапов чтения, сравнения и анализа.

Задача предварительного отбора различных последовательностей данных на основе относительного отклонения для формирования обучающих выборок в задачах машинного обучения формализуется следующим образом.

Требуется разбить последовательность файлов {F1, F2,…,FN} на непересекающиеся группы {G1 ,G2,…,GK}, такие что для любой пары соседних файлов Fi ,F i+1 ∈ G k выполняется условие:

$$\forall j \in \{1,\ldots,6\}: \frac{\max_t \left| x_{t,j}^{(i)} - x_{t,j}^{(i+1)} \right|}{\max_t \left| x_{t,j}^{(i+1)} \right|} \le \varepsilon, \quad (3)$$

где ε = 0,005 – заданная точность, а $x_{t,j}$ – значение j-го признака в файле.

Результатом работы алгоритма является список групп последовательных файлов, которые можно считать идентичными в пределах заданной точности. В дальнейшем из каждой такой группы будет сохранён только один файл, что позволит сформировать сокращённый, но репрезентативный набор данных для обучения нейронной сети без потери информативности и с существенным снижением вычислительных затрат.

Рассмотрим методы сравнения данных из традиционной практики предобработки. В машинном обучении универсальной формулой для количественной оценки различий является формула нормы разности двух векторов [7]:

$$\left| x^{(1)} - x^{(2)} \right|_p. \quad (4)$$

Однако при работе с нормированными концентрациями (где сумма компонент равна 1) и малыми значениями ($<10^{-6}$) абсолютные нормы становятся неинформативными, так как масштаб признаков сильно различается [8].

Для скалярных полей часто применяется относительная разность

$$\delta_j = \frac{\left| x_j^{(1)} - x_j^{(2)} \right|}{\max\left(\left| x_j^{(1)} \right|, \left| x_j^{(2)} \right| \right) + \varepsilon}, \quad (5)$$

где ε – малая константа для избегания деления на ноль. Такой подход широко используется в инженерной практике для оценки сходимости стационарных решений [9, 10].

Вместо сравнения самих значений, в некоторых исследованиях предлагается анализировать первую или вторую производную по времени, чтобы выявлять переходные процессы [11]. Однако для задач предобработки

обучающих данных это избыточно, достаточно выявить стационарные участки.

Итоговая архитектура обработки данных разделена на три независимых, но логически связанных этапа, каждый из которых оптимизирован для выполнения в пуле отдельных процессов (рисунок 1).

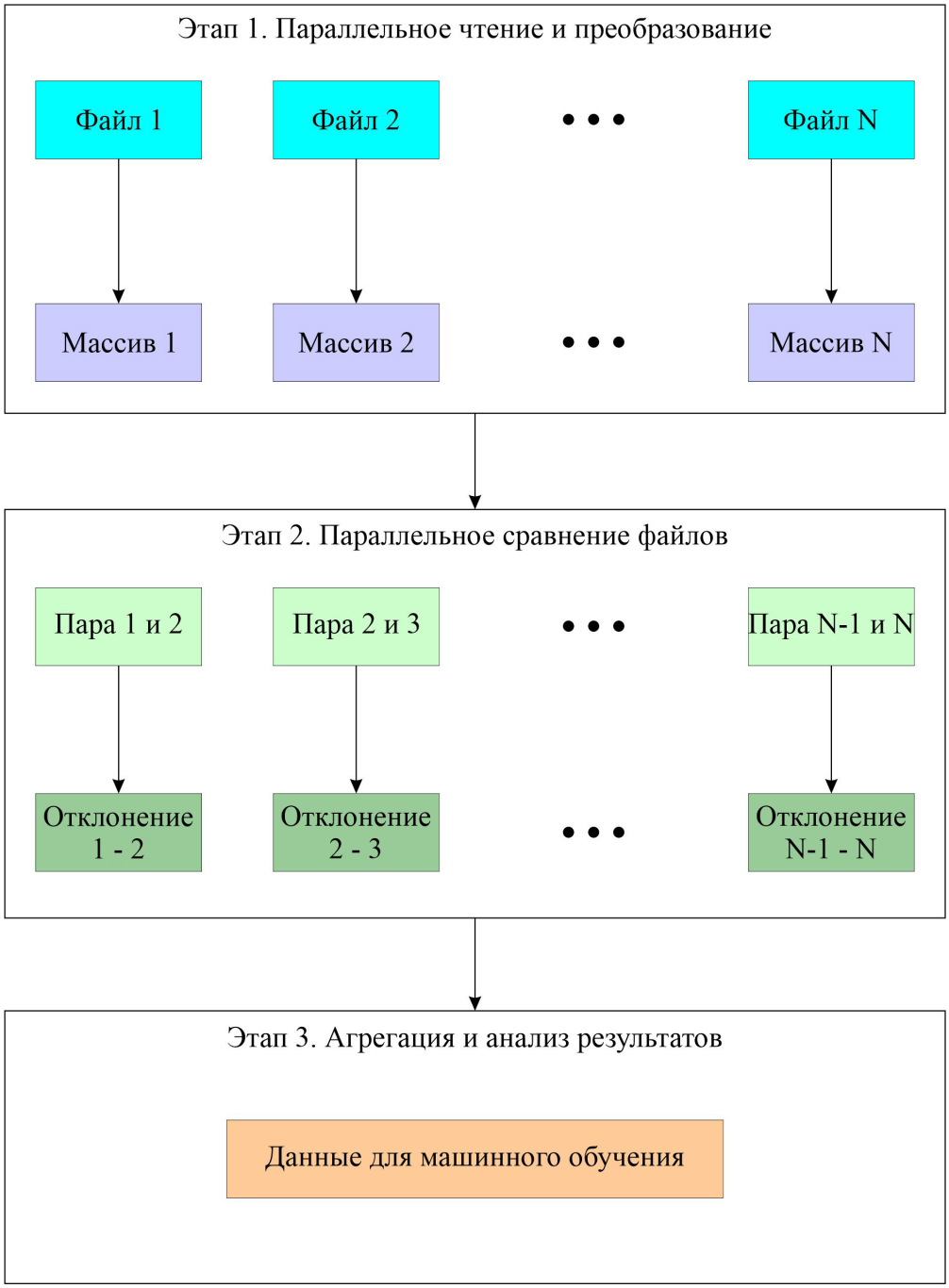

Рис. 1. – Архитектура обработки данных

На первом этапе осуществляется параллельное чтение исходных данных из файловой системы, где каждый файл загружается и преобразуется в компактный массив данных библиотеки NumPy в отдельном рабочем процессе. Это позволяет не только ускорить ввод-вывод, но и обеспечить отказоустойчивость, так как ошибки при чтении отдельных файлов локализуются и не прерывают обработку всего набора.

На втором этапе выполняется массовое сравнение соседних временных файлов системы. Сравнение происходит по формуле (5), с использованием константы $\varepsilon = 10^{-15}$. Каждая пара файлов обрабатывается независимо, что идеально соответствует парадигме «map» в параллельных вычислениях. Третий этап представляет собой запись полученных результатов в отчет и формирование набора данных для задачи машинного обучения.

Заключение

Методика предварительного отбора различных последовательностей данных на основе относительного отклонения для формирования обучающих выборок предназначена для решения важной проблемы машинного обучения – из обширного набора временных рядов или последовательных данных отобрать репрезентативное и разнообразное подмножество для обучения модели, чтобы она была устойчивой и хорошо описывала исследуемую задачу. Идея заключается в том, чтобы характеризовать каждую последовательность данных неким численным дескриптором, вычисляемым на основе относительного отклонения её точек от некоторого базового уровня (например, от среднего, тренда или предыдущего значения), а затем использовать этот дескриптор для стратифицированного отбора данных в обучающую выборку. Применение такой методики к задаче отбора данных для построения суррогатной модели численного моделирования процессов в нефтехимии показало хорошие результаты, позволяющей выделить стационарные режимы, не несущие информации для обучения нейросетевых

моделей. Это позволяет удалить неинформативные данные, повысить точность моделей машинного обучения, ускорить время обучения и избежать переобучения моделей.

Исследование выполнено за счет гранта Российского научного фонда № 25-21-00440, rscf.ru/project/25-21-00440/.

Литература

1. Solovev S.A., Soloveva O.V., Akhmetova I.G., Vankov Y.V., Paluku D.L. Numerical simulation of heat and mass transfer in an open-cell foam catalyst on example of the acetylene hydrogenation reaction // ChemEngineering. 2022. V. 6(1). article 11.

2. Галяутдинова А.Р., Ившин И.В., Соловьев С.А. Система оценки и прогнозирования технического состояния силового маслонаполненного трансформаторного оборудования распределительных сетей с применением машинного обучения // Известия Вузов. Проблемы энергетики. 2024. Т. 26, № 2. С. 32-45.

3. Киямов Р.Р., Мосева М.С. Обзор методов машинного обучения в задаче классификации водителей // Инженерный вестник Дона, 2024, № 7. URL: ivdon.ru/ru/magazine/archive/n7y2024/9370/.

4. Матренин П.В. Применение онтологического моделирования для автоматического отбора значимых признаков и семантической регуляризации моделей машинного обучения при разработке интеллектуальных информационных систем в электроэнергетике // Инженерный вестник Дона, 2025, № 9. URL: ivdon.ru/ru/magazine/archive/n9y2025/10368/.

5. Парасич В.А., Парасич И.В., Волович Г.И., Некрасов С.Г., Парасич А.В. Переобучение в машинном обучении: проблемы и решения // Вестник Южно-Уральского государственного университета. Серия:

Компьютерные технологии, управление, радиоэлектроника. 2024. Т. 24. №. 2. С. 18-27.

6. Добровольская Е.Д., Гундина М.А. Нормализация экспериментальных данных // Новые направления развития приборостроения: Материалы 16-й Международной научно-технической конференции молодых ученых и студентов. Минск: БНТУ, 2023. С. 228.

7. Lange M., Zühlke D., Holz O., Villmann T., Mittweida S.G. Applications of lp-Norms and their smooth approximations for gradient based learning vector quantization // ESANN. – 2014. – pp. 271-276.

8. Ferziger J.H., Perić M., Street R.L. Computational Methods for Fluid Dynamics. N-Y: Springer, 2020, 596 p.

9. Roache P.J. Verification and validation in computational science and engineering. Albuquerque, NM: Hermosa, 1998. Т. 895. P. 895.

10. Хусаинов Д.Я., Диблик Й., Кузьмич Е.И. Оценки сходимости решений линейного уравнения нейтрального типа // Динамические системы. 2006. Вып. 21. С. 43-53/

11. Karpov A.I., Kudrin A.V., Alies M.Y. Calculation of the stationary flame propagation velocity by the variational principle of irreversible thermodynamics // Case Studies in Thermal Engineering. 2022. V. 30. article 101767.

References

1. Solovev S.A., Soloveva O.V., Akhmetova I.G., Vankov Y.V., Paluku D.L. ChemEngineering. 2022. V. 6(1). Article 11.

2. Galyautdinova A.R., Ivshin I.I., Solovev S.A. Investiya Vuzov: Problemi Energetiki. 2024. V. 26, № 2. pp. 32-45.

3. Kiyamov R.R., Moseva M.S. Inzhenernyj vestnik Dona, 2024, № 7. URL: ivdon.ru/ru/magazine/archive/n7y2024/9370/.

4. Matrenin P.V. Inzhenernyj vestnik Dona, 2025, № 9. URL: ivdon.ru/ru/magazine/archive/n9y2025/10368/.

5. Parasich V.A., Parasich I.V., Volovich G.I., Nekrasov S.G., Parasich A.V. Vestnik Yuzhno-Ural'skogo gosudarstvennogo universiteta. Seriya: Komp'yuternye tekhnologii, upravlenie, radioelektronika. 2024. V. 24. №. 2. pp. 18-27.

6. Dobrovol'skaya E.D., Gundina M.A. Novye napravleniya razvitiya priborostroeniya: Materialy 16-j Mezhdunarodnoj nauchno-tekhnicheskoj konferencii molodyh uchenyh i studentov. Minsk: BNTU, 2023. P. 228.

7. Lange M., Zühlke D., Holz O., Villmann T., Mittweida S.G. ESANN. 2014. pp. 271-276.

8. Ferziger J.H., Perić M., Street R.L. Computational Methods for Fluid Dynamics. N-Y: Springer, 2020, 596 p.

9. Roache P.J. Verification and validation in computational science and engineering. Albuquerque, NM: Hermosa, 1998. V. 895. P. 895.

10. Husainov D.Ya., Diblik J., Kuz'mich E.I. Dinamicheskie sistemy. 2006. № 21. pp. 43-53.

11. Karpov A.I., Kudrin A.V., Alies M.Y. Case Studies in Thermal Engineering. 2022. V. 30. article 101767.

Авторы согласны на обработку и хранение персональных данных.

Дата поступления: 18.11.2025

Дата публикации: 25.12.2025