TECHNICAL SCIENCES

OPTIMIZATION OF RULE NUMBER IN A FUZZY DIGITAL TWIN Of A BIOREACTOR USING CLUSTERING TECHNIQUES

Mammadazada K.
Azerbaijan State Oil and Industry University,
BA Programs, Employee on Educational Technologies
http://orcid.org/0009-0007-7977-5122

ОПТИМИЗАЦИЯ КОЛИЧЕСТВА ПРАВИЛ В НЕЧЕТКОМ ЦИФРОВОМ ДВОЙНИКЕ БИОРЕАКТОРА С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ КЛАСТЕРИЗАЦИИ

Мамедзаде К.
Азербайджанский Государственный Университет Нефти и Промышленности,
BA Programs, Сотрудник по Образовательным Технологиям,
http://orcid.org/0009-0007-7977-5122
https://doi.org/10.5281/zenodo.17352739

Abstract

The development of accurate and computationally efficient digital twins for bioreactor systems remains a critical challenge in modern process engineering. This study presents a modern approach for optimizing the rule base of a fuzzy logic–based digital twin of a bioreactor through the application of clustering techniques. In the proposed method, clustering algorithms are employed to identify representative data patterns within the process dataset, allowing for the automatic generation and reduction of fuzzy rules without compromising model accuracy. Several clustering approaches, including fuzzy c-means (FCM) and subtractive clustering, are evaluated to determine their effectiveness in defining optimal rule numbers and membership functions. The resulting digital twin provides a more compact and interpretable model capable of accurately predicting bioreactor performance under varying operational conditions. Simulation results demonstrate that the clustering-based optimization significantly improves computational efficiency, reduces model complexity, and enhances generalization capability compared to traditional fuzzy modeling methods. This study contributes to the advancement of intelligent modeling strategies for bioprocess systems and supports the integration of digital twins in sustainable wastewater treatment applications.

Аннотация

Разработка точных и вычислительно эффективных цифровых двойников для биореакторных систем остаётся одной из преоритетных задач современной инженерии процессов. В данном исследовании представлен современный подход к оптимизации базы правил нечеткого цифрового двойника биореактора посредством применения методов кластеризации. В предложенном методе алгоритмы кластеризации используются для выявления репрезентативных шаблонов данных в наборе данных процесса, что позволяет автоматически формировать и сокращать количество нечетких правил без снижения точности модели. Рассмотрены различные подходы к кластеризации, включая метод нечетких с-средних (FCM) и субтрактивную кластеризацию, с целью оценки их эффективности в определении оптимального числа правил и функций принадлежности. Полученный цифровой двойник представляет собой более компактную и интерпретируемую модель, которая способна точно прогнозировать работу биореактора при различных эксплуатационных условиях. Результаты моделирования показывают, что оптимизация на основе кластеризации существенно повышает вычислительную эффективность, снижает сложность модели и улучшает её способность к обобщению по сравнению с традиционными методами нечеткого моделирования. Данная работа вносит вклад в развитие интеллектуальных стратегий моделирования биотехнологических процессов и способствует интеграции цифровых двойников в устойчивые системы очистки сточных вод.

Keywords: Fuzzy logic, clustering, bioreactor, wastewater treatment.
Ключевые слова: нечеткая логика, кластеризация, биореактор, очистка сточных вод.

ВВЕДЕНИЕ

В последние годы концепция цифровых двойников получила широкое распространение в области инженерии и управления технологическими процессами. Цифровой двойник является виртуальной моделью физической системы, воспроизводящее поведение в реальном времени и дает возможность анализа, прогнозирования и оптимизации процессов. В биотехнологической отрасли, в частности при управлении биореакторами, цифровые двойники играют ключевую роль в обеспечении устойчивости, энергоэффективности и качества процессов очистки сточных вод и биосинтеза.

Одним из перспективных подходов к построению цифровых двойников является использование нечеткой логики. Применение нечеткой логики позволяет моделировать сложные нелинейные зависимости и неопределенности, характерные для биопроцессов. Однако одной из основных проблем

нечетких моделей является избыточность базы правил. Следовательно большое количество правил усложняет модель, а это означает снижение интерпретируемости и увеличение вычислительных затрат. Поэтому оптимизация количества правил без потери точности является важной задачей для повышения эффективности нечетких цифровых двойников.

Поэтому методы кластеризации представляют собой эффективный инструмент для автоматического определения структуры базы правил. Кластеризационные алгоритмы, такие как fuzzy c-means (FCM) и субтрактивная кластеризация, позволяют выделить репрезентативные группы данных, на основе которых формируются оптимальные правила нечеткой логики. Применение этих методов обеспечивает баланс между точностью модели и её сложностью, а также способствует повышению вычислительной эффективности и способности модели к обобщению.

Целью данного исследования является разработка и анализ подхода к оптимизации числа правил в нечетком цифровом двойнике биореактора с использованием кластеризационных алгоритмов. Предложенный подход направлен на создание более компактной и интерпретируемой модели, способной точно описывать и прогнозировать поведение биореактора при различных эксплуатационных условиях, что имеет важное значение для внедрения интеллектуальных систем управления в биотехнологических процессах.

Цель

Целью данного исследования является разработка и обоснование подхода к оптимизации количества нечетких правил в цифровом двойнике биореактора с использованием методов кластеризации. Исследование направлена на повышение эффективности моделирования биотехнологического процесса за счёт сокращения числа правил базы знаний нечеткой модели при сохранении адекватности и точности описания динамики системы.

Для достижения этой цели предполагается проанализировать влияние структуры нечеткой модели на качество аппроксимации, исследовать различные методы кластеризации входных данных (в том числе Fuzzy C-Means, Gustafson–Kessel и Subtractive Clustering) для формирования оптимального количества и параметров нечетких правил, а также провести сравнительный анализ результатов моделирования до и после оптимизации. Предложенный подход позволяет сократить вычислительную нагрузку и время обучения нечеткой системы, что особенно важно при построении цифровых двойников биореакторов в режиме реального времени, без существенной потери точности моделирования динамических процессов.

Начальные условия

В данном исследование в качестве объекта моделирования рассматривается аэробный биореактор, используемый для очистки сточных вод. Модель разрабатывается на основе набора симуляционных данных, которые отражают функционирование установки в различных режимах работы. Для построения данного нечеткого цифрового двойника биореактора использовалась адаптивная нейро-нечеткая система вывода (ANFIS), которая основанна на обучении по экспериментальным данным, содержащим входные параметры $q_W, q_R, q_A, S_S, S_O, X_H$ и выходную переменную $\frac{dS}{dt}$.

В качестве начальных условий было принято, что структура ANFIS формируется автоматически с использованием методов кластеризации, которые определяют количество правил и параметры функций принадлежности. Количество нечетких правил варьировалось в диапазоне от 2 до 10. Данное количество правил позволило оценить влияние структуры базы правил на точность аппроксимации и обобщающую способность модели. Для исследования данные были разделены на две выборки: обучающая выборка, которая составляет 70%. Эти данные использовалась для построения модели. Следующя выборка это составляло 30% данных относится к Тестовая выборке. Тестовая выборка применялась для оценки точности и устойчивости модели к переобучению.

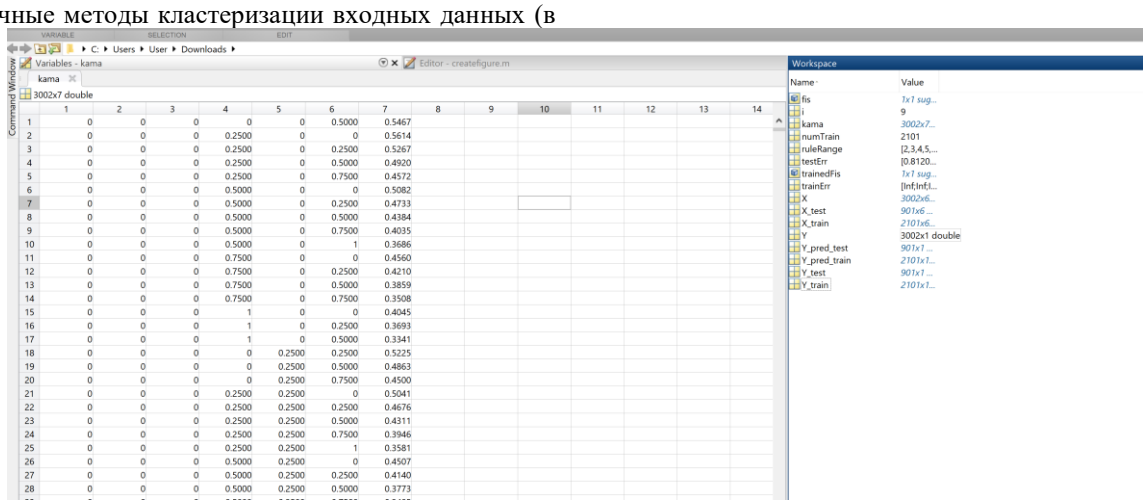

Рисунок 1.Матрица исходных данных и рабочее пространство MATLAB для моделирования ANFIS

На рисунке 1 показано рабочее окно MATLAB, который демострирует структуру исходных данных, использованных для построения нечеткой модели биореактора. Также показана таблица переменных, содержащая нормализованные значения входных и выходных параметров, а в правой части отображено содержимое рабочего пространства MATLAB (Workspace) с основными переменными моделирования.

Данные предварительно нормализованы в диапазоне [0, 1] для обеспечения корректной работы алгоритмов кластеризации и обучения ANFIS. Такой подход позволяет уравнять влияние всех переменных и избежать смещения результатов при анализе.

На основе рисунка 1 была выполнена кластеризация входных данных с целью автоматического определения количества нечетких правил в модели ANFIS. Для этого применялись алгоритмы типа Subtractive Clustering и Fuzzy C-Means (FCM), которые позволили выявить естественные группы (кластеры) в пространстве входных переменных. Каждому кластеру соответствует одно нечеткое правило, что делает выбор числа кластеров ключевым этапом оптимизации структуры модели.

Результаты кластеризации использовались для инициализации параметров функций принадлежности входных переменных и построения начальной структуры нечеткой системы. Далее выполнялось обучение ANFIS по обучающей выборке, а затем оценка качества по тестовой, что позволило определить оптимальное число правил, минимизирующее ошибку модели.

Методы

Методологическая основа данного исследования включает построение нечеткого цифрового двойника биореактора с использованием адаптивной нейро-нечеткой системы вывода (ANFIS) и также оптимизацию числа нечетких правил на основе кластеризационного анализа данных.

В качестве исходных данных применялись нормализованные экспериментальные и моделированные значения технологических параметров биореактора, которые включают расходы потоков q_W, q_R, q_A, концентрации растворённого субстрата S_S и кислорода S_O, а также концентрацию биомассы X_H. В качестве выходной переменной принята скорость изменения концентрации субстрата $\frac{dS}{dt}$, отражающая динамику протекающих биохимических процессов.

Для формирования начальной структуры нечеткой модели применялись методы кластеризации, позволяющие автоматически определить количество нечетких правил. В данном исследовании использовались алгоритмы Subtractive Clustering и Fuzzy C-Means (FCM), которые обеспечивают выявление естественных групп (кластеров) в пространстве входных данных. Количество выявленных кластеров соответствовало числу правил нечеткой системы, что позволило сократить размерность задачи и избежать субъективного выбора параметров модели.

После формирования структуры ANFIS модель проходила процесс обучения по обучающей выборке (70 % данных), а затем оценку точности по тестовой выборке (30 %). Для каждой структуры с различным количеством нечетких правил (от 2 до 10) вычислялись средние значения ошибок обучения и тестирования, выраженные в процентах.

Результаты данного анализа представлены на рисунке 2, где показана зависимость ошибки моделирования от количества нечетких правил ANFIS.

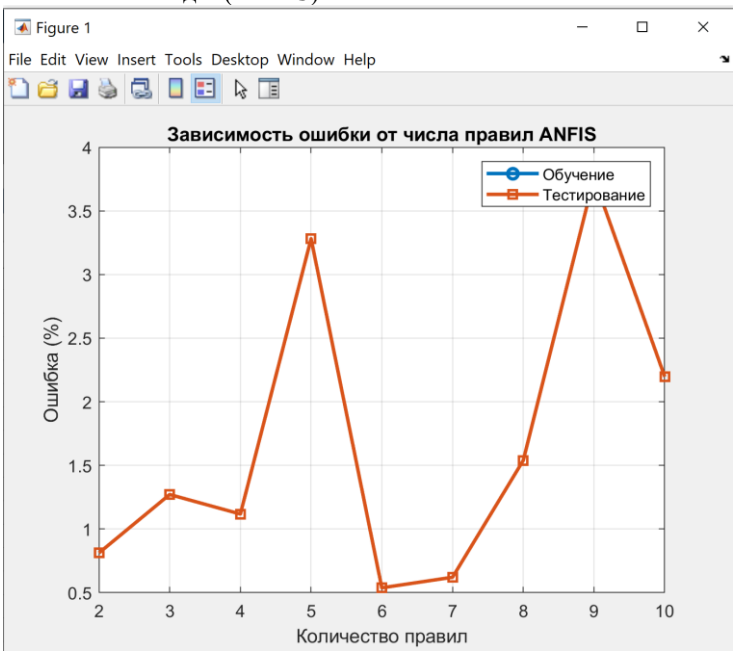

Рисунок 2. *Зависимость ошибки моделирования от количества нечетких правил ANFIS*

На рисунке 2 показан график зависимости средней относительной ошибки (%) от количества нечетких правил, определяющих структуру нечеткой модели ANFIS. Кривая с маркерами «Обучение» отражает значения ошибки, полученные на обучающей выборке. Из анализа представленных зависимостей видно, что при увеличении числа правил от 2 до 5 наблюдается умеренное изменение ошибки, однако при пяти правилах фиксируется значительный рост ошибки тестирования (до 3,3 %), что свидетельствует о неудачной структуре модели. При шести правилах достигается минимальная ошибка (около 0,5 %), что указывает на оптимальное соотношение между сложностью и точностью модели. При дальнейшем увеличении числа правил (8–10) ошибка вновь возрастает, что связано с эффектом переобучения и снижением обобщающей способности системы.

Таким образом, проведённый анализ позволяет сделать вывод, что оптимальное количество нечетких правил для построения цифрового двойника биореактора составляет шесть, что обеспечивает минимальную ошибку на тестовой выборке при сохранении высокой точности аппроксимации на обучающих данных.

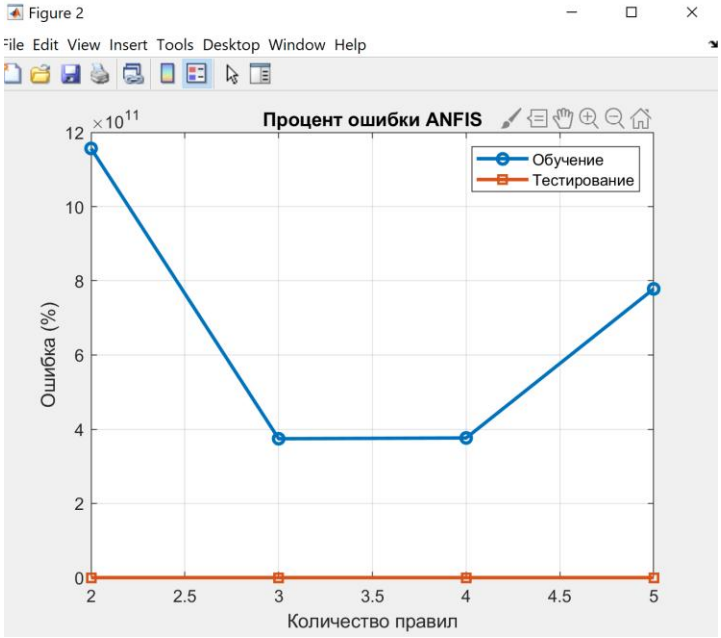

Рисунок 3. Зависимость процентной ошибки ANFIS от количества нечетких правил

На рисунке 3 показана зависимость процентной ошибки модели ANFIS от количества нечетких правил, которые формировались на этапе структурной идентификации цифрового двойника биореактора. Из графика видно, что при увеличении количества правил от двух до пяти ошибка обучения изменяется нелинейно. Минимальное значение ошибки наблюдается при трёх и четырёх правилах, где система демонстрирует наилучшее качество аппроксимации. При этом ошибка тестирования во всём диапазоне остаётся практически нулевой, что указывает на высокую устойчивость модели и отсутствие признаков переобучения.

Результаты, представленные на рисунке 3, оптимальное количество нечетких правил для рассматриваемого биореактора составляет три–четыре, что обеспечивает компромисс между точностью аппроксимации и вычислительной эффективностью ANFIS-модели.

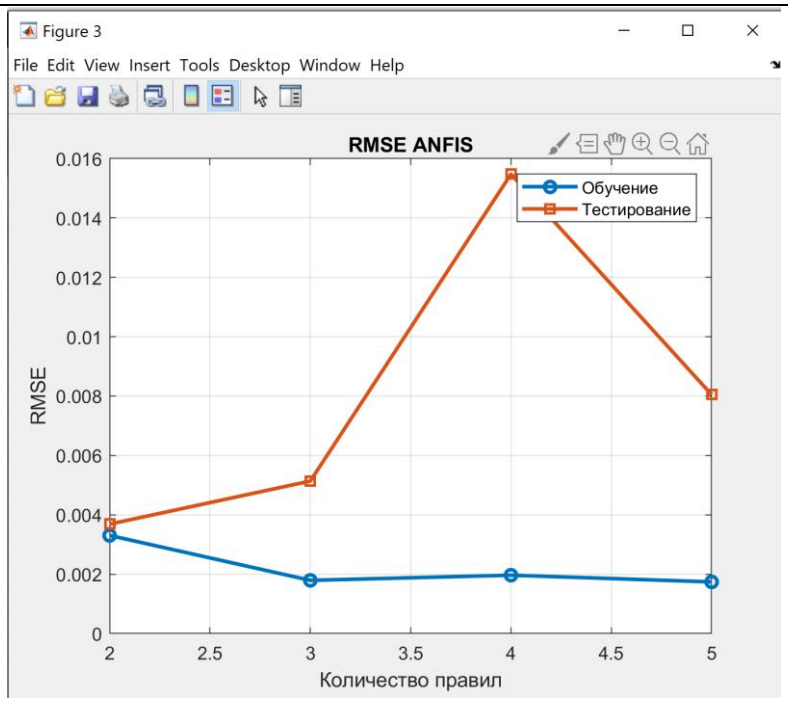

Рисунок 4. Зависимость среднеквадратической ошибки (RMSE) ANFIS от количества нечетких правил

На рисунке 4 представлена зависимость среднеквадратической ошибки (Root Mean Square Error, RMSE) модели ANFIS от количества нечетких правил, которые формируются при построении цифрового двойника биореактора. На графике показано количество нечетких правил и значение RMSE. Из анализа графика видно, что значения ошибки обучения остаются стабильно низкими (в пределах 0.0015–0.0035) во всём диапазоне рассматриваемых правил, что свидетельствует о корректной настройке параметров ANFIS и высокой точности аппроксимации обучающих данных. В то же время ошибка тестирования демонстрирует выраженный максимум при четырёх правилах (около 0.015), указывая на частичное переобучение модели. Минимальные значения RMSE для тестовой выборки наблюдаются при двух и пяти правилах, что соответствует более устойчивому обобщению данных.

Таким образом, по результатам представленные на рисунке 4, подтверждают, что при определении оптимального количества нечетких правил необходимо учитывать не только процентную ошибку, но и показатель RMSE как более чувствительный критерий, позволяющий оценить баланс между сложностью и точностью модели.

Представленные результаты позволили установить, что минимальные значения RMSE достигаются при количестве правил от двух до трёх, что соответствует наиболее сбалансированной конфигурации ANFIS-модели цифрового двойника биореактора.

Выводы

На основании сделанного анализа зависимостей ошибок обучения и тестирования ANFIS-модели можно выделить, что количество нечетких правил оказывает значительное влияние на точность и устойчивость модели цифрового двойника биореактора.

При малом числе правил (2–3) наблюдается недостаточная гибкость модели и умеренное недообучение, тогда как при чрезмерном количестве правил (8–10) проявляется эффект переобучения, выражающийся в росте тестовой ошибки и ухудшении обобщающей способности. Оптимальное количество нечетких правил определено в диапазоне от 3 до 6.

В этом диапазоне достигается минимальная тестовая ошибка и стабильные значения RMSE, что подтверждает эффективное представление нелинейных зависимостей между технологическими параметрами биореактора при минимальной вычислительной сложности модели. Результаты кластеризации входных данных подтвердили адекватность выбранного подхода к структурной идентификации ANFIS.

Применение алгоритмов Subtractive Clustering и Fuzzy C-Means позволило автоматически сформировать базу правил, исключив субъективный фактор при выборе их количества и параметров функций принадлежности. Анализ среднеквадратической ошибки (RMSE) показал высокую точность и устойчивость модели.

Низкие значения RMSE на обучающих данных (до 0.002) и умеренные на тестовых (до 0.015) свидетельствуют о корректной настройке параметров ANFIS и способности модели адекватно описывать динамику биореактора. Предложенный подход обеспечивает баланс между точностью и вычислительной эффективностью цифрового двойника.

Оптимизация структуры ANFIS по количеству правил позволяет сократить время обучения и снизить вычислительные затраты без потери качества моделирования, что делает данный метод перспективным для применения в задачах мониторинга и управления биотехнологическими процессами в реальном времени.

Declarations

The manuscript has not been submitted to any other journal or conference.

Study Limitations

There are no limitations that could affect the results of the study.

Список литературы:

1. Джанг Дж.-С. Р. ANFIS: адаптивная нечеткая система вывода на основе нейронной сети // *IEEE Transactions on Systems, Man, and Cybernetics*. – 1993. – Т. 23, № 3. – С. 665–685. DOI: 10.1109/21.256541

2. Такаги Т., Сугэно М. Идентификация нечетких систем и её применение к моделированию и управлению // *IEEE Transactions on Systems, Man, and Cybernetics*. – 1985. – Т. 15, № 1. – С. 116–132. DOI: 10.1109/TSMC.1985.6313399

3. Заде Л. А. Нечеткие множества // *Information and Control*. – 1965. – Т. 8, № 3. – С. 338–353. DOI: 10.1016/S0019-9958(65)90241-X

4. Шао Ц., Ван Л., Чжоу Ц. Цифровые двойники для моделирования и оптимизации биотехнологических процессов с использованием методов машинного обучения // *Computers & Chemical Engineering*. – 2021. – Т. 151. – Статья № 107347. DOI: 10.1016/j.compchemeng.2021.107347

5. Фати М., Кайнак О. Концепция цифрового двойника для киберфизических систем: обзор и анализ // *IEEE Access*. – 2018. – Т. 6. – С. 10883–10901.
DOI: 10.1109/ACCESS.2018.2800685.

6. Бездек Дж. С. Распознавание образов с использованием нечетких алгоритмов оптимизации. – Нью-Йорк: Springer, 1981. – 256 с. *(Классический источник по алгоритму нечеткой кластеризации FCM).*