

ПРОГНОЗИРОВАНИЕ КОНВЕРСИИ АЦЕТИЛЕНА ДЛЯ ПРОЦЕССА ГИДРОХЛОРИРОВАНИЯ НА БЕЗМЕТАЛЛИЧЕСКИХ КАТАЛИЗАТОРАХ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ, ОСНОВАННЫМИ НА ДЕРЕВЬЯХ РЕШЕНИЙ

Е.К. Кислинский¹, Д.В. Никитенко², А.Ю.Максимова¹

К.Д. Кобец², С.А. Корнев¹

¹ИПММ, г. Донецк

²ИНФОУ им. Л.М. Литвиненко, г. Донецк

gkislinskiy@gmail.com, nykytenko.dv@gmail.com, maximova.alexandra@mail.ru,

kiriloo.ru@mail.ru, sergkornev2001@mail.ru

Введение

Производство поливинилхлорида (ПВХ), одного из самых распространенных синтетических материалов, используемых в строительстве, медицине и производстве товаров повседневного спроса, связано с серьезными экологическими проблемами. Одним из промышленных способов производства сырья (винилхлорида мономера) для получения ПВХ является процесс гидрохлорирования ацетилена, для которого традиционным катализатором является хлорид ртути, нанесенный на активированный уголь. Такие катализаторы токсичны и приводят к загрязнению окружающей среды. В связи с этим поиск альтернативных каталитических систем, таких как безметаллические углеродные материалы, становится важной задачей. Ускорить процесс подбора таких катализаторов и снизить затраты на эксперименты можно применив методы машинного обучения (МО), а именно алгоритмы, основанные на деревьях решений.

Цель данного исследования – разработать метод предсказания активности безметаллических углеродных катализаторов в реакции гидрохлорирования ацетилена на основе методов МО, а также оценить важность факторов, влияющих на реакционный процесс.

Основная часть

Так как экспериментальные данные представлены в табличном виде и содержат до 1000 примеров, для экспериментов были выбраны усовершенствованный алгоритм градиентного бустинга от Yandex Catboost и алгоритм случайного леса [1–3]. Данные алгоритмы использовались для решения задач как классификации, так и регрессии. Анализ факторов проведен с помощью алгоритма Шепли [4].

База данных включает собранные из литературы экспериментальные данные [5], содержащие следующие параметры:

- Максимальная конверсия ацетилена (Max_Conv, %);
- Температура реакции (T, K);
- Объемная скорость подачи реакционной смеси (GHSV, ч⁻¹);

- Мольное соотношение $\text{HCl}:\text{C}_2\text{H}_2$;
- Удельная поверхность катализатора (S_{cat} , $\text{м}^2/\text{г}$);
- Элементный состав катализатора (содержание N, O, S, %);
- Масса загруженного катализатора (m_{cat} , г).

Исходная база данных состояла из 377 строк (систем). Однако из-за наличия пропущенных значений количество систем, пригодных для обучения, оказалось меньше. Для обеспечения максимальной точности исследования использовались только строки с полным набором данных. Обучение проводили на семи уникальных выборках, сформированных путем комбинирования различных наборов переменных. Перед использованием данные были очищены от пустых ячеек, результат представлен в таблице 1.

Таблица 1: Структура данных обучающих выборок.

№	Число строк	Наборы параметров для обучения
1	224	Max_Conv, GHSV, T, $\text{HCl}:\text{C}_2\text{H}_2$, S_{cat} , N%
2	224	Max_Conv, GHSV, T, $\text{HCl}:\text{C}_2\text{H}_2$, S_{cat} , N%, S%
3	140	Max_Conv, GHSV, T, $\text{HCl}:\text{C}_2\text{H}_2$, S_{cat} , N%, S%, 'O%
4	154	Max_Conv, GHSV, T, $\text{HCl}:\text{C}_2\text{H}_2$, N%, m_{cat}
5	122	Max_Conv, GHSV, T, $\text{HCl}:\text{C}_2\text{H}_2$, S_{cat} , N%, m_{cat}
6	122	Max_Conv, GHSV, T, $\text{HCl}:\text{C}_2\text{H}_2$, S_{cat} , N%, S%, m_{cat}
7	91	Max_Conv, GHSV, T, $\text{HCl}:\text{C}_2\text{H}_2$, S_{cat} , N%, S%, O%, m_{cat}

Первый этап анализа данных предполагал трансформацию задачи регрессии в задачу классификации. С этой целью диапазон значений максимальной конверсии ацетилена, варьирующийся от 0 до 100%, был разделен на 10 равных интервалов (децилей), каждый из которых был обозначен как отдельный класс. Такой подход не только упрощает интерпретацию результатов, но также минимизирует влияние выбросов в данных, что особенно важно при работе с неоднородными или зашумленными наборами. Кроме того, классификация на основе децилей позволяет более четко выделить закономерности и тенденции, которые могут быть скрыты при анализе непрерывных значений. Это особенно полезно для задач, где точное прогнозирование численных значений менее критично, чем общее понимание распределения и группировки данных.

Для решения задачи классификации были выбраны два алгоритма: CatBoostClassifier и RandomForestClassifier. Эти методы были выбраны по следующим причинам:

- CatBoostClassifier: этот алгоритм эффективно работает с категориальными переменными, автоматически обрабатывает пропущенные данные и устойчив к переобучению благодаря механизму упорядоченного бустинга.
- RandomForestClassifier: этот метод прост в интерпретации, устойчив к выбросам и позволяет оценивать важность признаков, что делает его полезным для анализа данных.

Второй подход заключался в предсказании значений максимальной конверсии в качестве непрерывной переменной. Для этой задачи были выбраны модели CatBoostRegressor и RandomForestRegressor.

Для оценки производительности моделей использовались метрики: ROC-AUC (оценивает способность модели различать классы, что важно для задач классификации с несбалансированными данными), MAE (средняя абсолютная ошибка, измеряет среднюю величину ошибок в задачах регрессии), R² (коэффициент детерминации, показывает долю дисперсии, объяснённую моделью) и RMSE (среднеквадратическая ошибка, учитывает крупные отклонения за счёт квадратичного штрафа).

Для моделей классификации была получена высокая предсказательная способность: значения ROC-AUC лежат в диапазоне [0.73; 0.80].

В свою очередь, регрессионные алгоритмы случайного леса способны давать лучшие предсказания максимальной конверсии ацетилена, чем CatBoost. Для регрессионных моделей значения R² лежат в диапазоне [0.85; 0.97], в то время как для бустинга эти величины составляют [0.61; 0.74]. Значения метрик MAE и RMSE (оценка на тестовых данных) у случайного леса значительно меньше, чем у бустинга.

Для определения ключевых факторов, влияющих на максимальную конверсию ацетилена, был проведен анализ важности признаков с использованием метода SHAP [4]. Наибольшее влияние на величину максимальной конверсии ацетилена оказывают следующие параметры (перечислены в порядке убывания их значимости):

1. Количество азота в углеродном носителе (N%) – этот параметр оказывает наибольшее влияние. Каталитически активными центрами являются атомы азота, находящиеся в структуре углеродного катализатора. Чем больше каталитических центров, тем выше конверсия.
2. Температура реакции (T) – температура играет ключевую роль в кинетике реакции – скорость реакции растёт с повышением температуры.
3. Скорость потока реакционной смеси (GHSV) – этот параметр влияет на время контакта реагентов с катализатором, что напрямую связано с конверсией. Чем выше скорость потока, тем меньше время контакта реагентов с катализатором и как следствие меньше конверсия.

4. Удельная поверхность катализатора (S_{cat}) – чем больше удельная поверхность, тем больше площадь контакта реагентов с катализатором, что также важно для эффективности реакции.

Заключение

Разработаны и протестированы модели усовершенствованного алгоритма градиентного бустинга от Yandex Catboost и алгоритм случайного леса для предсказания максимальной конверсии ацетилена в реакции газофазного гидрохлорирования ацетилена. Полученные результаты позволяют сформулировать рекомендации по оптимизации процесса газофазного гидрохлорирования ацетилена: для регулирования величины максимальной конверсии ацетилена стоит сфокусироваться на основных ключевых параметрах, таких как содержание азота в катализаторе, температура реакции, скорость потока и удельная поверхность катализатора. Это может способствовать повышению эффективности процесса и снижению затрат.

Работа выполнена при финансовой поддержке Минобрнауки России в рамках научной темы «Разработка и совершенствование интеллектуальных методов классификации и прогнозирования для задач распознавания образов и моделирования информационных процессов» FREM-2024-0001 (Регистрационный номер 1023111000141-9-1.2.1).

Список литературы:

1. Просиз Дж. Прикладное машинное обучение и искусственный интеллект для инженеров – Астана: Изд-во «АЛИСТ», 2024. 424 с. ISBN 978-60109-5051-1.
2. CatBoost – Текст: электронный // Хабр: официальный сайт. 2023. URL: <https://habr.com/ru/companies/lanit/articles/778714/> (дата обращения: 10.02.2025).
3. Scikit-learn: Machine Learning in Python // scikit-learn : [сайт]. 2024. URL: <https://scikit-learn.org/stable/index.html> (дата обращения: 10.02.2025). – Текст: электронный.
4. SHAP (SHapley Additive exPlanations) : Explain Any Machine Learning Model // SHAP Documentation : [сайт]. 2024. URL: <https://shap.readthedocs.io/en/latest/index.html> (дата обращения: 11.02.2025). – Текст: электронный.
5. Qiao X., Zhao Z.-H., Zhang J. Progress in mercury-free catalysts for acetylene hydrochlorination // Catalysis Science and Technology. 2024. Vol. 14. Issue 14. Pp. 3838–3852. <https://doi.org/10.1039/d4cy00549j>