УДК 004.89
AMS MSC2020: 68Q45

Комплекс алгоритмов Data Mining в исследовании процесса протекания химических реакций

Биллиг В. А., Звягинцев Н. В.

Тверской государственный технический университет

Аннотация. В настоящее время накоплено значительное количество экспериментальных данных, фиксирующих процесс протекания химических реакций. Анализ этих данных комплексом алгоритмов Data Mining дает важную практическую информацию для поиска эффективных условий проведения реакций, при которых получается максимальное количество целевого продукта при минимальных затратах. В данной работе на примере работы с базой, содержащей данные о протекании реакции карбонилирования различных олефинов, показано, как разработанный нами программный комплекс, позволяет извлечь полезные знания, способствующие повышению эффективности протекания химических реакций.

Ключевые слова: Data Mining, Apriori, деревья решений, химические реакции.

Введение

Химические реакции (ХР) являются сложными процессами, на протекание которых оказывает влияние масса условий: давление, температура, состав и природа взаимодействующих веществ и катализаторов. Вместе с тем, важной задачей является поиск наиболее эффективных условий, когда при минимальных затратах получается максимальное количество целевого продукта. Для решения этой задачи необходимо проанализировать влияние всех условий на протекание ХР. Решение можно осуществить несколькими путями — моделированием механизма и кинетики самой ХР [2], а также применением комплекса алгоритмов Data Mining к экспериментальным

данным о протекании ХР. Второй подход чаще всего является менее затратным с вычислительной точки зрения, однако требует существенного объема экспериментальных данных и разработки подходов к кодированию информации о факторах протекания химических реакций [1].

1. Формирование БД

Важным этапом второго подхода является формирование БД с результатами экспериментов при различных условиях. Данные об условиях протекания химических реакций являются слабосвязанными данными. В публикациях рассматривают влияние широкого набора факторов, данные могут быть указаны в разных единицах измерения. В рамках данной работы исследуется процесс протекания реакции карбонилирования олефинов. Данная ХР хорошо экспериментально изучена, поэтому удалось собрать достаточно представительную выборку, содержащую данные о влиянии различных условий на получение целевого продукта. В состав результирующей выборки включены:

- конверсия исходного вещества и селективность (для формирования выходных данных — целевые параметры),
- давление, температура, состав исходного вещества и состав катализатора (для входных параметров).

Данные о давлении были пересчитаны в атмосферах, данные о температуре приведены к градусам Цельсия, а на основе информации о химическом составе были сформированы атрибуты булева типа:

- l_has_p — в состав лиганда входит фосфор,
- substr_n_c — количество атомов углерода в исходном веществе,
- substr_ol — исходное вещество является спиртом,
- prec_cl — катализатор содержит хлор,
- acid_type — тип кислоты (органическая или не органическая).

Созданная выборка, содержащая 183 записи, является частью БД, содержащей больше входных параметров. Мы оставили наиболее информативные параметры, определенные по данным предварительных исследований.

2. Комплексный анализ данных

Эксперт, занимающийся анализом данных, обычно, хочет получить ответы на вопросы: «Каковы ожидаемые значения целевых параметров при заданном наборе входных параметров, каковы наиболее информативные параметры, влияющие на результат, и как управлять ими для достижения нужного выхода»? Зачастую для эксперта важны не столько конкретные значения параметров, сколько качественные оценки. Ему необходимы правила, позволяющие предсказать класс, которому принадлежит целевой параметр при управлении входными данными. Понятно, что эксперту нужны не только правила, но и характеристики этих правил, такие как достоверность правила, частота применения правила, возможно и другие характеристики. По нашему глубокому убеждению, ответы на эти вопросы можно получить, применяя широкий арсенал методов извлечения знаний из данных, известных как методы Data Mining.

3. Программная платформа

Для анализа данных мы используем программный комплекс, содержащий три модуля:

- Предварительная подготовка данных.
- Комплекс алгоритмов анализа данных.
- Визуализация и объяснение результатов анализа.

Для части алгоритмов анализа данных разработана собственная реализация, для части используются алгоритмы, входящие в пакет sklearn языка Python. Предварительная подготовка данных является важной частью работы с данными. На этом этапе решаются такие проблемы как восстановление пропусков в записях базы данных, а, главное, преобразование данных к виду, требуемому тем

или иным алгоритмом. Одни алгоритмы могут работать только с транзакционными данными, другие — с категориальными данными, третьи — с непрерывными данными, возможно, приведенными к одному масштабу. Комплекс алгоритмов включает алгоритмы:

- Построения ассоциативных правил.
- Различные вариации алгоритма кластеризации kmeans для работы с непрерывными данными и алгоритма CLOPE для работы с категориальными (транзакционными) данными.
- Построения деревьев решений для задачи классификации.
- Построения деревьев решений для задачи регрессии.

Остановимся на некоторых алгоритмах и результатах их работы.

4. Алгоритм ConApriori построения ассоциативных правил

У разработанного нами алгоритма ConApriori есть два существенных отличия в сравнении с классическим алгоритмом Apriori:

- способ представления данных,
- способ построения достоверных ассоциативных правил.

Каждая запись базы данных представлена одним числом. Числовое представление записей базы данных не зависит от размера записи и позволяет эффективно вычислять значение базовой функции Support за время $O(N)$ с минимальной константой, где N — число записей базы данных. В отличие от классического алгоритма, в котором строятся частые правила на основе ранее построенных частых правил, в данном алгоритме строятся достоверные правила на основе ранее построенных достоверных правил, что повышает эффективность алгоритма. Подробное описание алгоритма приведено в работе [3]. Алгоритм позволяет находить правила с заданной частотой и достоверностью. Кроме этого, для каждого правила вычисляется характеристика, называемая lift, определяющая степень корреляции между посылкой правила и его заключением. Приведем

несколько правил, найденных в результате работы алгоритма для исследуемой БД:

$P2 \Rightarrow sel\_tar1$: частота = 0,21;
достоверность = 0,93; лифт = 2,54

$P2, T1 \Rightarrow sel\_tar1$: частота = 0,21;
достоверность = 0,93; лифт = 2,54

$P2, acid\_type3 \Rightarrow sel\_tar1$: частота = 0,21;
достоверность = 0,95; лифт = 2,60

5. Алгоритм DecisionTreeClassifier построения дерева классификации

Этот алгоритм, реализованный в пакете tree, входящем в состав пакета sclearn, для построения дерева классификации использует два критерия — примесь Джини и энтропию. Примесь Джини считается по следующей формуле:

$$gini = \sum_{i=1}^{N}(1 - p_i^2) \tag{1}$$

Для критерия энтропии применяется известная формула Шеннона:

$$entropy = -\sum_{i=1}^{N} p_i * \log_2(p_i) \tag{2}$$

В обеих формулах используются p_i — вероятности появления классов, рассчитываемые как частоты появления класса в выборке. Для наших данных оба критерия строят качественно похожие деревья. Приведем правила, которые можно вывести из анализа дерева решений. Дерево решений для классификации целевого параметра sel_tar строится по обучающей выборке, содержащей 128 записей, в которых целевой параметр представлен тремя классами примерно равной мощности. В корне дерева имеет место следующая ситуация: $(gini = 0.66, samples128(48, 33, 47))$. При построении дерева находятся два наиболее информативных параметра — давление и температура. При выполнении условий: $if(P > 52.2 \ \& \ T < 105)$ уже на 3-м

шаге приходим в узел ($gini = 0.37, samples34(1, 7, 26)$), где уже можно применять правило с примерной частотой 0.25 и достоверностью 0.75, о том, что целевой параметр sel_tar принадлежит третьему классу. При выполнении условия: $if(P >= 52.2 \quad \& \quad T < 72.5)$ приходим в узел (узел ($gini = 0.24, samples37(32, 1, 4)$). Здесь решение о том, что целевой параметр принадлежит к первому классу принимается примерно с той же частотой, но с более высокой достоверностью, приближающейся к 0.9. Эти данные хорошо согласуются с приведенными выше ассоциативными правилами.

Размер тезисов не позволяет даже кратко характеризовать другие используемые алгоритмы анализа данных.

Заключение

1) Рассматриваемая нами БД позволяет извлечь знания, которые могут помочь эксперту, занимающемуся исследованием эффективности протекания процесса карбонилирования олефинов.
2) Комплексное применение алгоритмов Data Mining повышает надежность и способствует улучшению качества анализа данных.

Список литературы

[1] *Биллиг, В. А.* Информационная система обработки и хранения данных о кинетике химических реакций / В. А. Биллиг, Н. В. Звягинцев // Программные продукты и системы. — 2018. — Т. 31, №. 4. — С. 808–813.

[2] Исследование влияния природы лигандов на региоселективность реакции карбонилирования стирола в присутствии комплексов палладия (II) / Н. В. Звягинцев, О. Л. Елисеев, Л. Т. Кондратьев, А. Л. Лапидус // Доклады Академии наук. — 2010. — Т. 434, №. 2. — С. 189–195.

[3] *Billig, V. A.* Effective algorithm for constructing associative rules // Программные продукты и системы. — 2017. — Т. 30, №. 2. — С. 196–206.

Библиографическая ссылка

Биллиг, В. А. Комплекс алгоритмов Data Mining в исследовании процесса протекания химических реакций / В. А. Биллиг, Н. В. Звягинцев // Всероссийская научная конференция «Математические основы информатики и информационно-коммуникационных систем». Сборник трудов. — Тверь : ТвГУ, 2021. — С. 118–124.
https://doi.org/10.26456/mfcsics-21-19

Сведения об авторах

1. Владимир Арнольдович Биллиг
Тверской государственный технический университет. Профессор

Россия, Тверь, наб. Афанасия Никитина, 22
E-mail: vladimir-billig@yandex.ru

2. Николай Васильевич Звягинцев
Тверской государственный технический университет. Аспирант

Россия, Тверь, наб. Афанасия Никитина, 22
E-mail: n.zvyagintsev@gmail.com