

# СПЕКТРАЛЬНЫЙ АНАЛИЗ ВЕЩЕСТВ НА ОСНОВЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ

Д. О. Захаревич

*Белорусский государственный университет, г. Минск;*

*zahdanik@gmail.com;*

*науч. рук. – Д. В. Щегрикович, канд. физ.-мат. наук, доц.*

В работе рассмотрено решение задачи обработки большого объема данных при анализе спектров для определения состава веществ. Для достижения максимальной эффективности проанализирована работа нескольких алгоритмов машинного обучения, решающих задачу регрессии. Создана модель для определения состава сенажа злаковых трав по спектру в ближнем инфракрасном свете. Полученный алгоритм интегрирован в приложение для удобства использования.

**Ключевые слова:** спектральный анализ; регрессия; машинное обучение; дерево решений; случайный лес; SVM; гребневая регрессия; LASSO; ElasticNet.

## ВВЕДЕНИЕ

В современной науке и технике, для того чтобы определить химический состав веществ, используют множество различных методов. Минералы, найденные геологами, и новые вещества, полученные химиками, характеризуются, прежде всего, по составу. Для правильного ведения технологических процессов в разных отраслях необходимо точное знание химического состава данного сырья. Химические методы анализа не всегда соответствуют требованиям техники и науки. В связи с этим на практику внедряются физико-химические и физические методы исследования, которые являются более точными. Среди этих методов одно из значимых мест занимает спектральный анализ, имеющий множество ценностей и преимуществ. [1]

Спектральный анализ широко используется в различных отраслях промышленности и науки, и служит универсальным инструментом, который позволяет точно и оперативно исследовать элементный состав вещества. Эта информация необходима для правильного ведения технологических процессов, контроля качества исходных материалов, промежуточного и готового продуктов, а также позволяет создавать новые материалы с заданными качествами. [2]

## ПОСТАНОВКА ЗАДАЧИ

Проблема спектрального анализа заключается в необходимости обработки большого объема данных, получаемых в результате

измерения спектров исследуемых веществ. Таким образом ставится задача регрессионного анализа спектра и определение состава вещества на его основе. Для достижения максимальной скорости обработки и достоверности получаемых результатов необходимо проанализировать различные методы машинного обучения и выбрать наиболее эффективный.

## СРАВНЕНИЕ АЛГОРИТМОВ

Данные предоставлены лабораторией skarb-lab для совместного исследования образцов. Измерения получены с помощью спектрометра в ближнем инфракрасном диапазоне. Рассматривается спектр поглощения. Инфракрасная область представляет собой часть электромагнитного спектра, соседствующего с видимым диапазоном, включающее длины волн в диапазоне от 0.8 до 1000 мкм. Данная область обладает различными свойствами в зависимости от того, какой диапазон рассматривается. Так, например, в ближнем инфракрасном диапазоне наблюдаются широкие полосы поглощения.

Чтобы понять основы инфракрасной спектроскопии, нужно обратиться к квантово-механической интерпретации молекул. Молекула воды может проявлять симметричные и асимметричные колебания и/или изгиб в своем молекулярном дипольном моменте. Взаимодействие между инфракрасным излучением и молекулой может происходить только в том случае, если частота инфракрасного излучения и молекулярного дипольного момента совпадают. [3]

Полученные данные проходят предобработку. Записи с недостающими значениями удаляются, так как при использовании большого набора данных игнорирование малой доли кортежей не окажет существенного влияния на дальнейший анализ. Таким же образом проводится удаление строк, имеющих выбросы. Следующим шагом предобработки данных – нормализация для корректировки весов значений, подаваемых на вход модели. Затем данные разделяются на обучающую и тестовую выборки.

Для решения задачи регрессии созданы несколько моделей основанные на следующих методах: SVM, дерево решений, случайный лес, гребневая регрессия, LASSO, ElasticNet. Качество моделей определяется с помощью коэффициента детерминации  $R^2$ .

Результат применения исследуемых моделей на исходных данных отображен на рисунке 1.

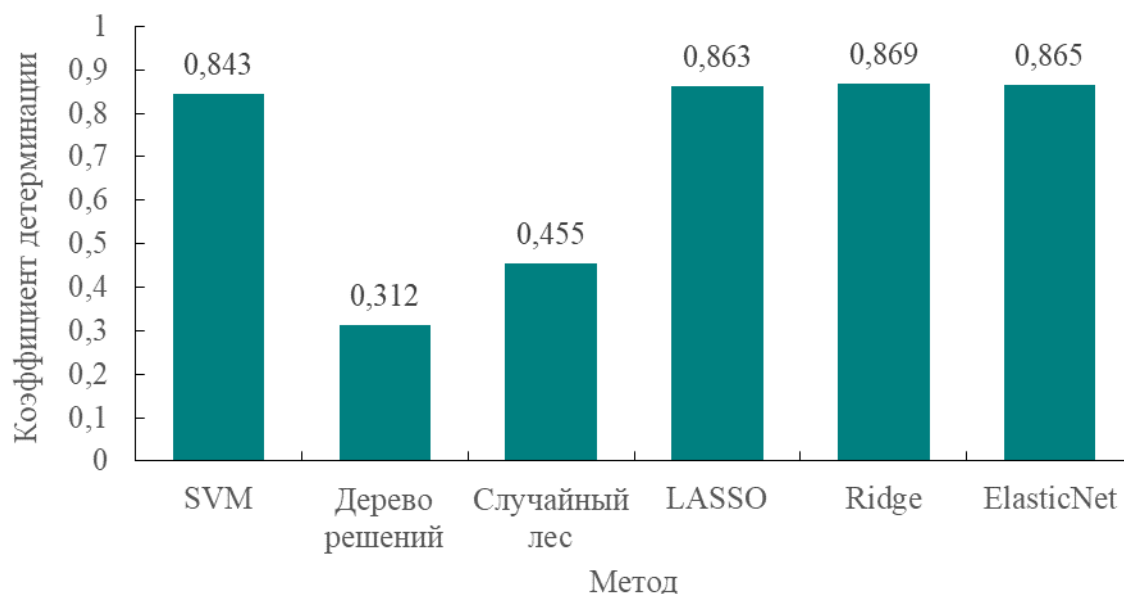


Рис. 1. Качество методов решения задачи регрессии оцененной с помощью коэффициента детерминации

## ПРИМЕНЕНИЕ RIDGE

Исходя из полученных оценок выбран метод гребневой регрессии (Ridge). Гребневая регрессия или ридж-регрессия (англ. ridge regression) — один из методов понижения размерности. Применяется для борьбы с избыточностью данных, когда независимые переменные коррелируют друг с другом, вследствие чего проявляется неустойчивость оценок коэффициентов многомерной линейной регрессии [4].

Качество работы модели дополнительно исследовано с помощью среднеквадратической ошибки (MSE). Она работает следующим образом: для каждой точки вычисляется разность квадратов между прогнозами и целью, после чего эти значения усредняются.

Значения MSE для некоторых параметров представлено в таблице. Среднее значение среднеквадратической ошибки составило 0.122

| Параметр         | MSE   | Параметр  | MSE   |
|------------------|-------|-----------|-------|
| pH               | 0.107 | Клетчатка | 0.056 |
| Уксусная кислота | 0.440 | Сахар     | 0.040 |
| Молочная кислота | 0.137 | Хлорин    | 0.210 |
| Протеин          | 0.028 | Лизин     | 0.048 |
| Жиры             | 0.074 | Метионин  | 0.075 |

## СОЗДАНИЕ ПРИЛОЖЕНИЯ

Разработанный алгоритм интегрируется в десктопное приложение, представляющее собой форму для заполнения необходимой информации и выбора файла спектра. На выходе создается pdf-файл содержащий результаты анализа.

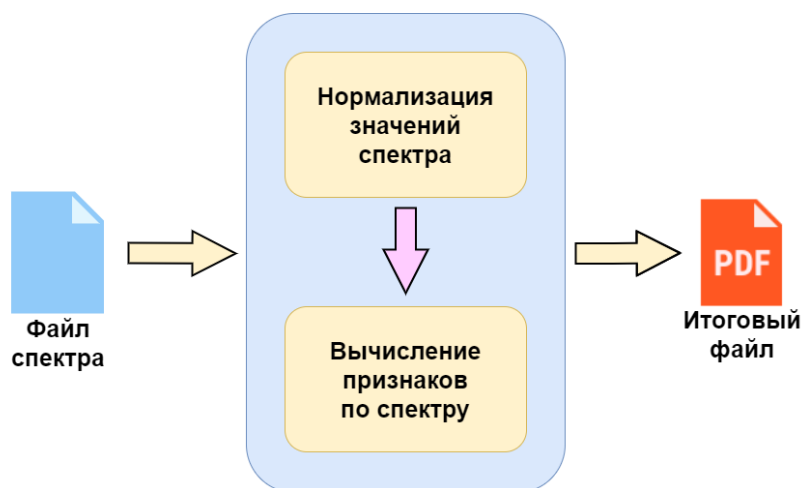


Рис. 2. Схема работы приложения

## ЗАКЛЮЧЕНИЕ

В результате работы проанализирован ряд алгоритмов машинного обучения (SVM, дерево решений, случайный лес, гребневая регрессия, LASSO, ElasticNet), решающих задачу регрессии. Наилучший результат показал метод гребневой регрессии (значение коэффициента детерминации  $R^2$  составило 0.87). На основе этого алгоритма создана модель для расчета состава сенажа злаковых трав по спектру в ближнем инфракрасном диапазоне. Полученная модель интегрирована в приложение.

### Библиографические ссылки

1. Агаркова Е. И. Спектральный анализ и его применение при проведении экспертных исследований // Студенческий научный форум, 2015.
2. Интернет-адрес: <https://www.iskroline.ru/analysis/primenenie-spektralnogo-analiza/>.
3. Интернет-адрес: <https://q-interline.com/home/technology/spectroscopy/>.
4. Интернет-адрес: [https://neerc.ifmo.ru/wiki/index.php?title=Вариации\\_регрессии](https://neerc.ifmo.ru/wiki/index.php?title=Вариации_регрессии).