Министерство науки и высшего образования Российской Федерации
Санкт-Петербургский политехнический университет Петра Великого
Физико-механический институт
Высшая школа теоретической механики и математической физики

Работа допущена к защите
Директор ВШТМиМФ
д.ф.-м.н., чл.-корр. РАН
_____ А. М. Кривцов
«___»_____20__ г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА БАКАЛАВРА

ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРЕДСКАЗАНИЯ ВЛИЯНИЯ НЕБОЛЬШИХ МОЛЕКУЛ НА ЭКСПРЕССИЮ ГЕНОВ

По направлению подготовки
01.03.03 «Механика и математическое моделирование»
Профиль
01.03.03_01 «Механика и математическое моделирование сред с микроструктурой»

Выполнил
студент гр.5030103/00101 Г. А. Сазыкин

Руководитель
Доцент ВШТМиМФ, к.т.н. В. Р. Мешков

Консультант
Ассистент ВШТМиМФ А. Д. Ершов

Санкт-Петербург
2024

САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО

Физико-механический институт

Высшая школа теоретической механики и математической физики

УТВЕРЖДАЮ

Директор ВШТМиМФ

А. М. Кривцов

«__»_____20__г.

ЗАДАНИЕ

на выполнение выпускной квалификационной работы

студенту Сазыкину Георгию Андреевичу, гр. 5030103/00101

1. Тема работы: Применение методов машинного обучения для предсказания влияния небольших молекул на экспрессию генов

2. Срок сдачи студентом законченной работы: 30.05.2024

3. Исходные данные по работе: набор данных об одноклеточных возмущениях мононуклеарных клеток периферической крови человека, алгоритмы визуализации многомерных данных, метод дисперсионного анализа, метод главных компонент, алгоритмы поиска гиперпараметров, кластеризации, регрессии.

4. Содержание работы (перечень подлежащих разработке вопросов): предобработка предоставленных данных и извлечение новых данных из свойств молекул, описательных статистик экспрессии генов; анализ с помощью методов визуализаций многомерных данных (t-SNE, UMAP); применение и анализ методов уменьшения размерности; масштабирование данных и кодирование категориальных переменных; сравнение между собой различных алгоритмов регрессии.

5. Перечень графического материала (с указанием обязательных чертежей): не предусмотрено

6. Консультанты по работе: Ершов А. Д., ассистент ВШТМиМФ

7. Дата выдачи задания 26.02.2024

Руководитель ВКР _____ Мешков В. Р., доцент ВШТМиМФ, к.т.н.

Задание принял к исполнению 26.02.2024

Студент _____ Сазыкин Г.А.

РЕФЕРАТ

На 41 с., 17 рисунков, 5 таблиц

ЭКСПРЕССИЯ ГЕНОВ, МАШИННОЕ ОБУЧЕНИЕ, МЕТОДЫ СНИЖЕНИЯ РАЗМЕРНОСТИ, СТАТИСТИЧЕСКИЕ ТЕСТЫ, МЕТОД ГЛАВНЫХ КОМПОНЕНТ, РЕГРЕССИЯ, ГРАДИЕНТНЫЙ БУСТИНГ

В данной работе проводится исследование данных об экспрессии генов в отдельных клетках периферической крови человека. Проведен разведочный анализ данных; проведено сравнение методов снижения размерности и визуализации, а также статистических критериев сравнения групп, описано решение задачи регрессии для предсказания экспрессии генов на тестовых данных и сравнение различных методов. Описан процесс извлечения новых данных из свойств молекул, масштабирования и кодирования категориальных переменных, подбора гиперпараметров для алгоритма градиентного бустинга.

THE ABSTRACT

41 pages, 17 pictures, 5 tables

GENE EXPRESSION, MACHINE LEARNING, DIMENSIONALITY REDUCTION METHODS, STATISTICAL TESTS, PRINCIPAL COMPONENT METHOD, REGRESSION, GRADIENT BOOSTING

In this paper, data on gene expression in individual cells of human peripheral blood are studied. An exploratory analysis of the data was carried out; methods of dimensionality reduction and visualization were compared, as well as statistical criteria for comparing groups, a solution to the regression problem for predicting gene expression on test data was described and various methods were compared. The process of extracting new data from the properties of molecules, scaling and encoding categorical variables, and selecting hyperparameters for the gradient boosting algorithm is describe

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Достижения в области одноклеточных технологий позволили получить представление о функционировании клеток и тканей на уровне ДНК, РНК, белков. Одним из таких примеров является технология одноклеточного секвенирования, которая позволяет получить данные о последовательности ДНК отдельной клетки, а также визуализировать пространственное положение нуклеиновых кислот [10]. Однако использование одноклеточных методов для разработки лекарств является трудоемким и дорогостоящим; при этом не все клетки и ткани поддаются высокопроизводительному транскриптомному скринингу. Повлиять на скорость и расширение разработки новых лекарственных препаратов путем прогнозирования химических изменений в различных клетках помогло бы применение методов математической статистики и методов машинного и глубокого обучения. Для прогнозирования реакции на лекарственные препараты было разработано несколько методов, в их числе, группа методов Dr.VAE, PertVAE, SSVAE – алгоритмы, основанные на вариационном автоэнкодере. Однако этим методам не хватает надлежащего набора данных для сравнительного анализа с различными типами клеток. Одним из самых больших доступных наборов является карта связи Cmap - библиотека содержит более 1.5 млн профилей экспрессии генов из 5000 низкомолекулярных соединений и 3000 генетических реагентов, протестированных на нескольких типах клеток [13]. Несмотря на это, Cmap включает наблюдения только 978 генов, что составляет около 5% от общего количества.

В ходе данной работы были поставлены следующие задачи:

1. Провести разведочный анализ данных.
2. Извлечь новые признаки.
3. Провести оценку данных с помощью статистических тестов.
4. Сравнить результаты работы различных алгоритмов регрессии для предсказания экспрессии для новых типов клеток и лекарственных препаратов.

ГЛАВА 1. РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ

1.1 Постановка эксперимента

Для работы с исследуемым набором данных важно знать постановку эксперимента, в результате которого была вычислена экспрессия генов. На 96-луночные планшеты размораживали и наносили клетки периферической крови человека, которые состоят из Т-клеток, В-клеток, NK-клеток, миелоидные клеток. Две колонки планшета предназначены для исследования реакции на дабрафениб и белиностат, потому что эти лекарственные препараты способны оказывать значительное влияние на транскрипцию; данная часть эксперимента называется положительным контролем. Другой столбец предназначен для отрицательного контроля, иначе оценивается реакция на ДМСО – вещество выступает в качестве растворителя для соединений, которые использовались в данном исследовании. Остальные лунки на планшете отведены каждому из 72 соединений. Полный набор данный включает в себя 2 планшета с различными соединениями на каждого из 3-х доноров, всего 6 планшетов. В каждой из лунок не всегда можно получить информацию обо всех типах представленных клеток, потому что на некоторые из них соединения могут оказывать токсичное воздействие.

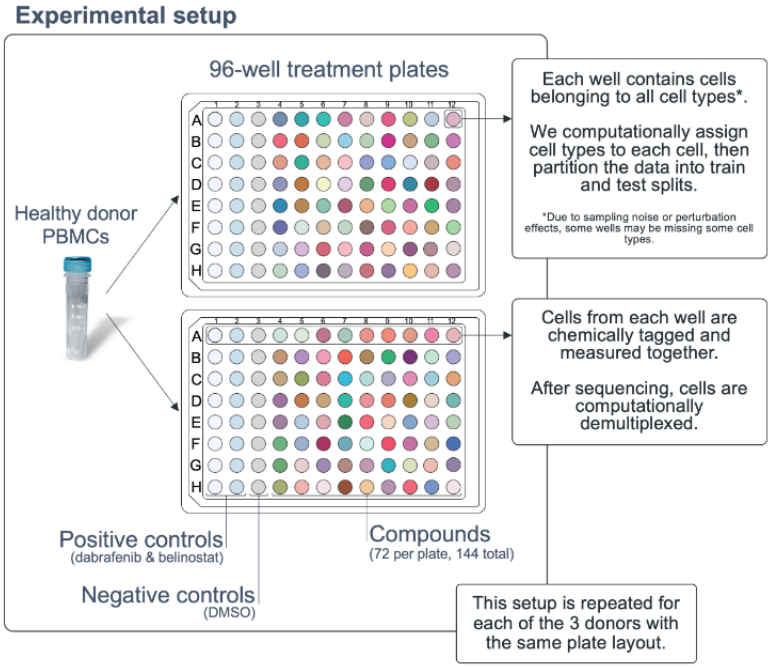

Рис. 1.1 Технические подробности эксперимента

В данной постановке задачи необходимо смоделировать дифференциальную экспрессию миелоидных и B-клеток для большинства из заданных соединений. В качестве обучающего датасета выступают данные по 144 лекарственным соединениям, которые воздействовали на T-клетки (CD4+, CD8+, регуляторные клетки), NK- клетки; также включена информация о влиянии 10% соединений от общего количества на миелоидные и B-клетки, что позволит оценить влияние экспериментального изменения на уровень экспрессии для каждого из генов в транскрипции - в предоставленном наборе данных таких генов 18211. На рис. 1.2 отображено, каким образом разделяется набор данных.

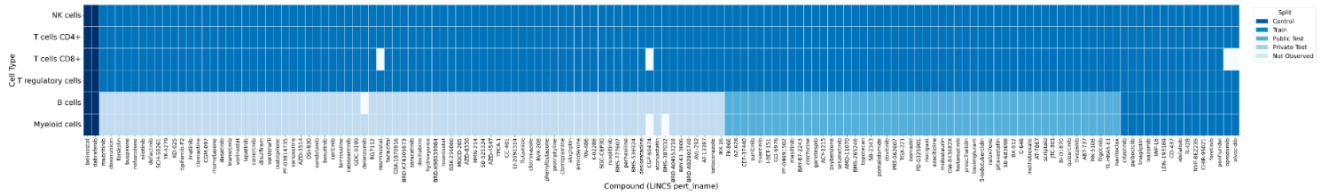

Рис. 1.2 Разделение тренировочной и тестовой выборки

Для оценки влияния каждого соединения на экспрессию необходимо усреднить исходные показатели экспрессии генов в каждой клетке определенного типа в каждом образце, что называется псевдообъемным расчетом [14]. С помощью программного пакета для языка программирования R Limma, по данным псевдообъемного расчета строится линейная модель, куда в качестве ковариат включаются технические переменные, описывающие донора крови, планшет, строку планшета и экспериментальная переменная, описывающая лекарственное соединение [6]. На рис 1.3 представлен принцип работы программного пакета Limma.

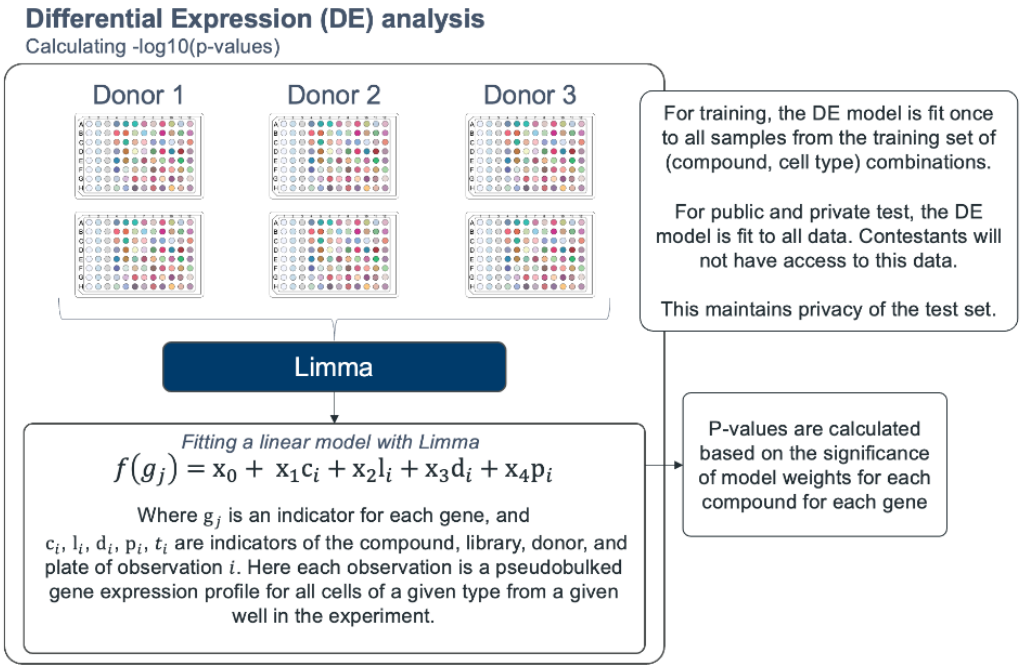

Рис. 1.3 Принцип работы Limma

Линейная модель позволит оценить кратное изменение экспрессии гена и получить скорректированное при множественном сравнении p-value, которое показывает, что экспрессия гена зависит от составной экспериментальной переменной; иначе - уровень значимости коэффициента в линейной модели перед переменной, отвечающей за лекарственное соединение.

1.2 Структура входных данных

В таблице 1.1 представлен набор полей входных данных.

Название переменной	Описание
Гены A1BG, A1BG-AS1, …, ZZEF1	значение дифференциальной экспрессии $(-\log_{10}(\text{p} - \text{value})*\text{sign(LFC)})$ для каждого гена. Здесь предполагаемое логарифмическое изменение экспрессии между обработанным и контрольным состоянием после

	усадки, рассчитанной с помощью пакета Limma. Положительный LFC означает, что уровень экспрессии гена повышается по сравнению с контролем. Всего 18211 значений
cell_type	аннотированный тип каждой клетки
sm_name	название лекарственного соединения
sm_lincs_id	Глобальный идентификатор соединения в библиотеке LINCS
SMILES	строка, представляющая из себя молекулу соединения, которое было использовано в эксперименте
control	логическое значение, указывающее, использовался ли данный экземпляр в качестве контроля

Таблица 1.1 Описание полей датасета

При решении задачи регрессии, проведения статистических тестов важно учитывать, какое распределение имеет целевая переменная, есть ли в исследуемых данных аномалии и выбросы. В качестве целевой переменной выступает средняя экспрессия по всем генам, присутствующих в датасете. Также стоит обратить внимание на то, соблюдается ли баланс классов – это важно для предсказательной способности моделей регрессии. На рис. 1.4 видно, что наблюдается дисбаланс классов; в тренировочном датасете наблюдений, связанных с клетками (B клетки, миелоидные клетки) меньше, чем наблюдений по остальным клеткам. В тестовом же датасете содержится информация только по B-клеткам и миелоидным клеткам. Вероятно, предсказательная способность модели регрессии на NK, T клетках будет лучше, чем на B и миелоидных клетках.

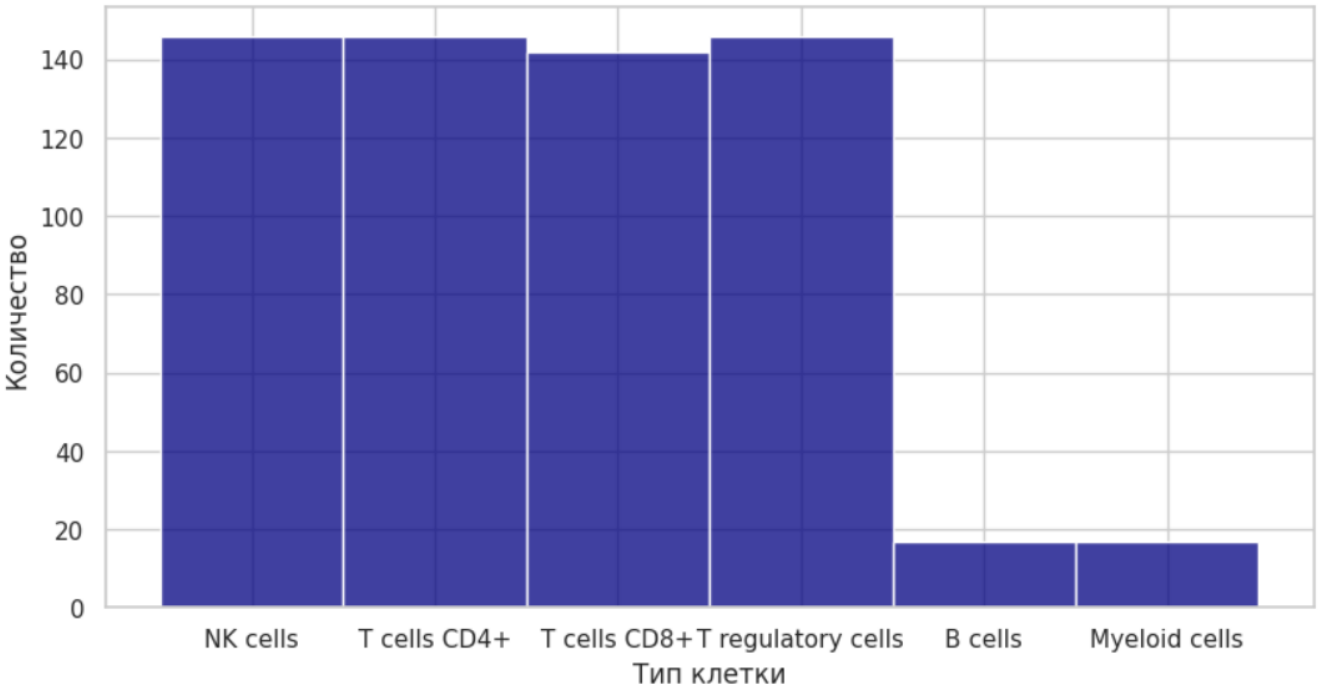

Рис. 1.4 Распределение количества наблюдений в каждой из клеток

Интерес также представляет гистограмма распределения средней экспрессии как по типу клеток (рис 1.5), так и в зависимости от того, являлась ли часть эксперимента положительным контролем. Вид графиков напоминает распределение Гаусса; но узнать, была ли взята выборка из нормального распределения, можно с помощью критерия Шапиро-Уилка.

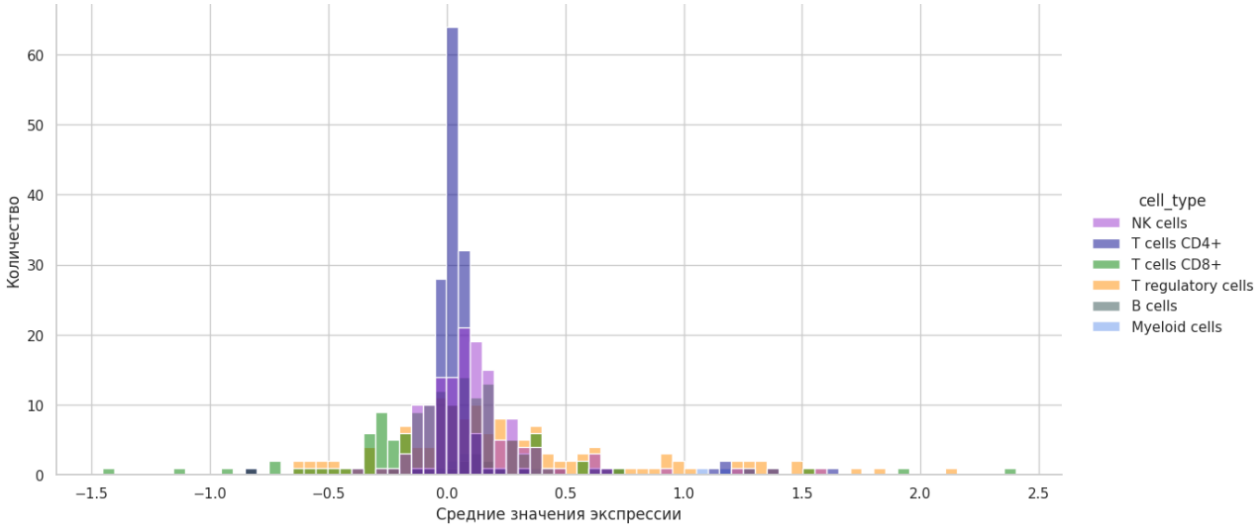

Рис. 1.5 Распределение средней экспрессии в зависимости от типа клеток

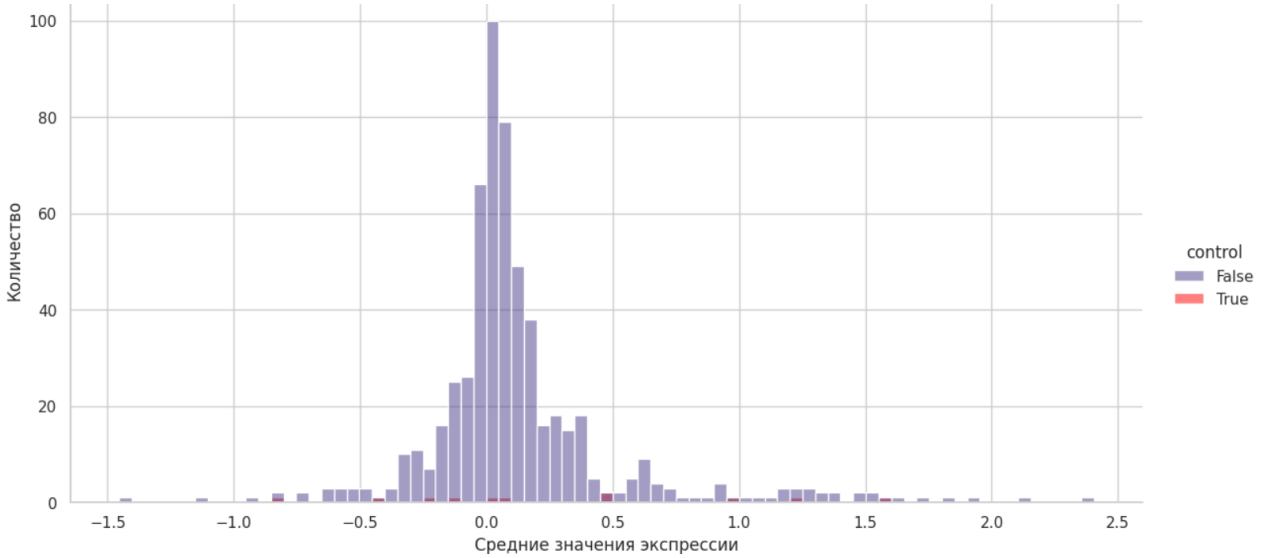

Рис. 1.6 Распределение средней экспрессии в зависимости от наличия положительного контроля

Нулевая гипотеза теста Шапиро-Уилка заключается в том, что случайная величина, выборка которой известна, распределена по нормальному закону. Альтернативная гипотеза: закон распределения не является нормальным. Критерий основан на отношении оптимальной линейной несмещенной оценки дисперсии к ее обычной оценке методом максимального правдоподобия [2]. Статистика критерия имеет вид:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad (1)$$

где $x_{(i)}$ – статистика i-го порядка (упорядоченные значения выборки), $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ – выборочное среднее, $(a_1, a_2, \dots a_n)$ – коэффициенты, значения которых можно узнать в соответствующей критерию таблице для заданного размера выборки n и порядкового номера .

Статистики критерия и значения p-value в результате тестирования приведены в таблице 1.2.

Тип клетки	W	p-value
NK	0.36	$1.29*10^{-22}$
B	0.56	$4.14*10^{-6}$

T CD4+	0.29	$1.22*10^{-23}$
T CD8+	0.74	$1.78*10^{-14}$
T регуляторные клетки	0.41	$8.19*10^{-22}$
Миелоидные клетки	0.67	$4.83*10^{-5}$

Таблица. 1.2 Результаты теста Шапиро-Уилка

По результатам теста для каждого из типа клеток мы можем отклонить нулевую гипотезу о том, что выборки были взяты из нормального распределения.

На рис 1.7 отображены графики QQ-plot –графики, позволяющие сравнить наблюдаемые квантили с квантилями нормального распределения. Во всех случаях значения сильно отклоняются от диагональной прямой - можно сделать вывод, что представленные выборки были взяты не из нормального распределения.

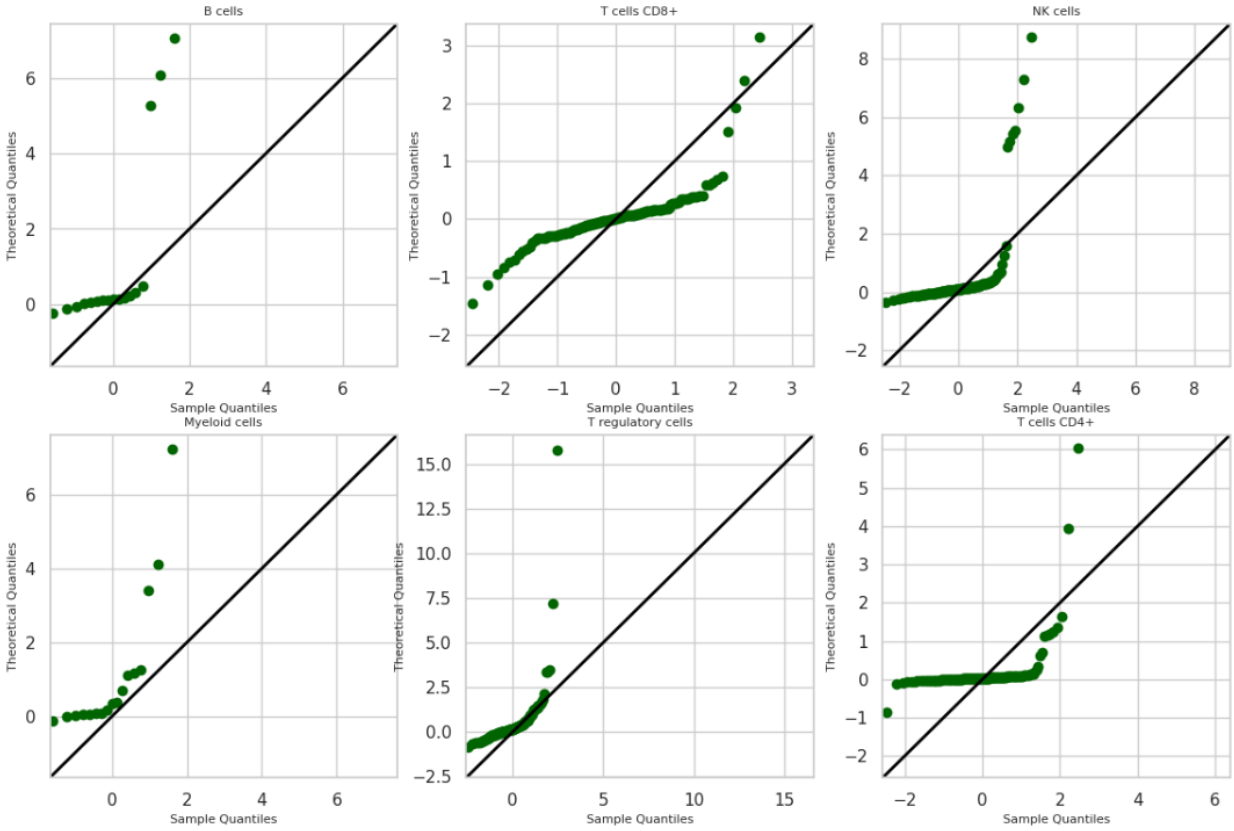

Рис. 1.7 QQ-plot для каждой из выборок по типам клетки

О распределении данных и наличии выбросов в каждой из групп можно узнать, совместив скрипичную диаграмму с диаграммой boxplot (рис 1.8). Во всех

группах наблюдается довольно много выбросов; тем не менее большинство значений средней экспрессии в каждой из групп сконцентрировано около нуля.

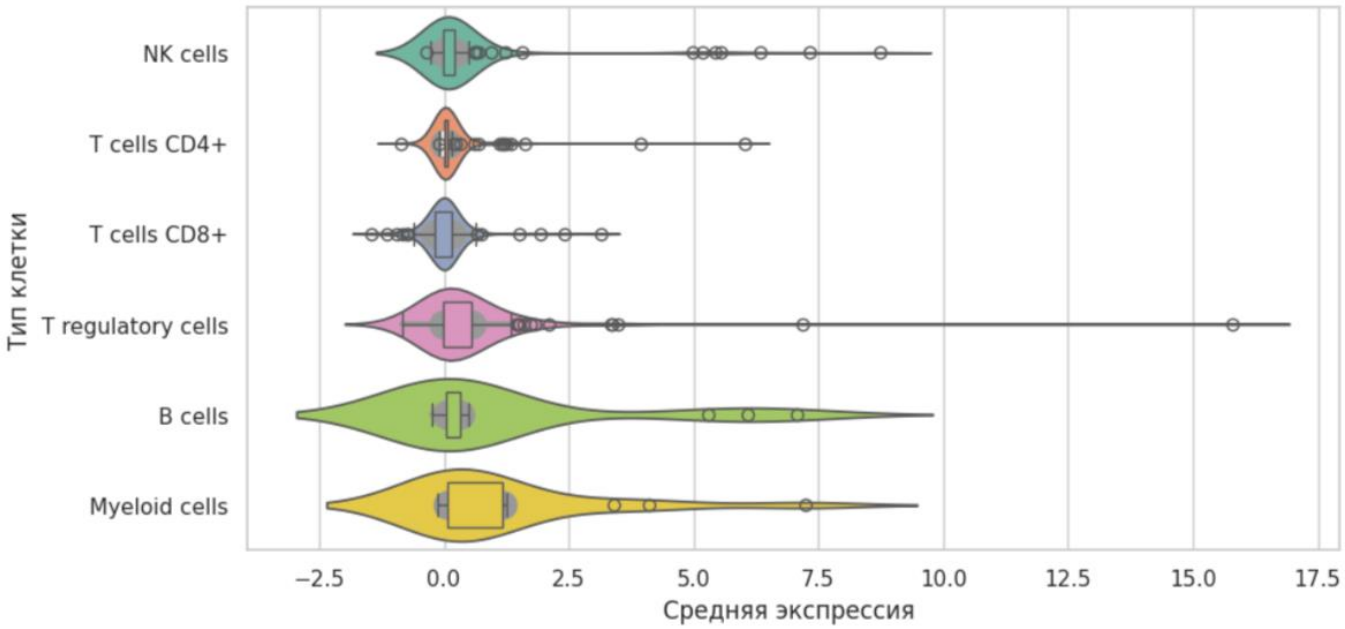

Рис. 1.8 Скрипичная диаграмма распределения средней экспрессии в зависимости от типа клетки

1.3. Проектирование признаков

В исходном наборе данных интерес, прежде всего, представляют две категориальные переменные: лекарственный препарат (sm_name) и тип клетки (cell_type). Переменная control не так важна для анализа в данной задаче: на рис 1.6 можно наблюдать, что наблюдается явный дисбаланс классов. Поле sm_lincs_id является технической переменной, которая связана с соединением. Переменная SMILES (упрощенная система молекулярного ввода) представляет из себя линейную строку символов в одноименном формате – такое представление используется для описания и представления молекулярных структур; в данном случае признак описывает молекулу лекарственного соединения. Чтение и работу с такими строка реализует программный пакет RDKit для языка программирования Python и C++, предназначенный для химоифнорматики и машинного обучения в биоинформатике. Библиотека RDKit также позволяет по строке визуализировать молекулу; пример такой визуализации отображен на рис. 1.9

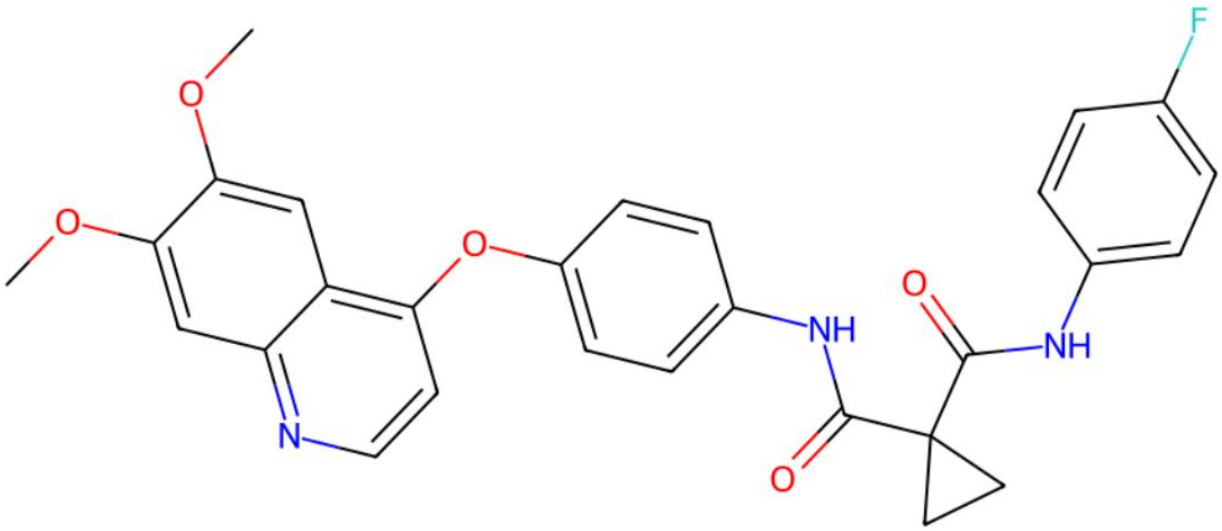

Рис. 1.9 Визуализация молекулы кабозантиниба с помощью средств библиотеки RDKit

По известной структуре соединения из молекулы можно получить информацию о количестве атомов, связей между ними; логарифм молярной рефракции, который является мерой поляризуемости одного моля соединения элемента. Новые признаки в дополнение к исходным данным формируются не только посредством получения информации о свойствах соединения. Сгруппировав исходные данные по типу клетки и по типу лекарственного соединения, можно найти среднее и дисперсию экспрессии генов – получившиеся колонки включаются в тренировочный и тестовый набор данных.

Некоторые из признаков представляют из себя категориальные переменные, которые необходимо привести в числовой вид- сделать это с помощью метода One Hot Encoding – алгоритм создает новые столбцы, в которых указывается, присутствует ли значение (значение “1”), или отсутствует (значение “0”). В таком случае количество новых столбцов будет равно количеству уникальных значений признака.

Переменные в получившимся наборе данных зачастую не имеют одинаковую размерность, что может повлиять на работу алгоритмов регрессии, основанные на кластеризации и поиске ближайших соседей по наблюдениям – возникает необходимость привести количественные признаки к заданному диапазону. В качестве такого преобразователя используется MinMaxScaler –

метод линейно масштабирует до фиксированного диапазона, но не уменьшает влияние выбросов. Преобразование определяется формулой

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}, \quad (2)$$

где X – вектор числовых значений, X_{min} – минимальное значение в выборке, X_{max} – максимальное значение выборки, X_{scaled} – вектор числовых значений, приведенный к заданному диапазону.

ГЛАВА 2. АЛГОРИТМЫ СНИЖЕНИЯ РАЗМЕРНОСТИ

2.1. Алгоритм t-SNE для решения задачи снижения размерности

В описанной выше постановке задачи имеется датасет об экспрессии генов; наблюдения описываются многомерной переменной с размерностью пространства 18211 – именно столько генов в исходном наборе данных. Чтобы выяснить, насколько наблюдения похожи между собой, необходимо получить новую переменную, существующую, например, в двумерном пространстве, которая бы в максимальной степени сохраняла бы структуру и закономерности в исходных данных. Постановка задачи имеет следующий вид:

- Пусть $x_1, x_2, \dots x_n$ – исходные числовые признаки
- Требуется найти $y_1, y_2, \dots y_d$, где d ≤ n - новые числовые признаки, такие, что при переходе к которым будет потеряно наименьшее количество исходной информации, а также будут наиболее корректно отображать данные в пространстве меньшей размерности.

Одним из методов, позволяющий решить задачу снижения размерности данных, является t-SNE (t-stochastic neighbor embedding). Алгоритм стохастического вложения соседей начинается с перехода многомерного евклидового расстояния между наблюдениями в условные вероятности. Объекты в исходном пространстве локально считаются нормально распределенными и близость между двумя объектами считается с использованием нормированной плотности нормального распределения – расстояние между объектами i и j вычисляется по формуле, которую можно интерпретировать, как вероятность того, что точка x_i выберет в качестве своего соседа точку x_j среди остальных точек данных [11]:

$$p(i|j) = \frac{\exp\left(-\frac{\left||x_i - x_j|\right|^2}{2\sigma_j^2}\right)}{\sum_{k \neq j} \exp\left(-\frac{\left||x_k - x_j|\right|^2}{2\sigma_j^2}\right)}, \quad (3)$$

где σ_j – стандартное отклонение переменной x_j.

Формула несимметрична относительно перестановки i и j, поэтому итоговое расстояние вычисляется как

$$p = \frac{p(i|j) + p(j|i)}{2} \quad (4).$$

В низкоразмерном пространстве сложно сохранить расстояние между объектами, если в пространстве большой размерности некоторые из объектов находились на небольшом расстоянии друг от друга. Сходство наблюдений в новом пространстве можно описывать с помощью распределения Коши, потому что за увеличение расстояния между объектами штраф будет небольшим.

Определим вероятности $q(i|j)$ для пространства низкой размерности:

$$q(i|j) = \frac{\left(1 + \left||y_i - y_j|\right|^2\right)^{-1}}{\sum_{k \neq j}\left(1 + \left||y_k - y_j|\right|^2\right)^{-1}}, \quad (5)$$

где $y_i,\ y_j$ – точки в пространстве низкой размерности.

Задача метода SNE – уменьшить разницу в распределении вероятностей. Таким образом, задача сводится к поиску таких координат y_j, чтобы расстояния между объектами в исходном пространстве и расстояния между объектами в пространстве проекций были похожи.

Функцией потерь является дивергенция Кульбака-Лейблера – мера для измерения различия вероятностей [11]:

$$KL(p||q) = \sum_{i \neq j} p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right) \rightarrow min \quad (6)$$

Выполнить нелинейное снижение размерности данных можно с помощью инструментов библиотеки sklearn языка программирования Python. Для сравнения было обучено 4 алгоритма t-SNE с количеством компонент, равным 2. Кроме того, был задан гиперпараметр perplexity который отвечает за нахождение σ_j с помощью вещественного бинарного поиска; параметр принимает значение 5, 15, 30, 40 соответственно. Результат работы алгоритмов отображен на рис. 2.1. и рис 2.2.

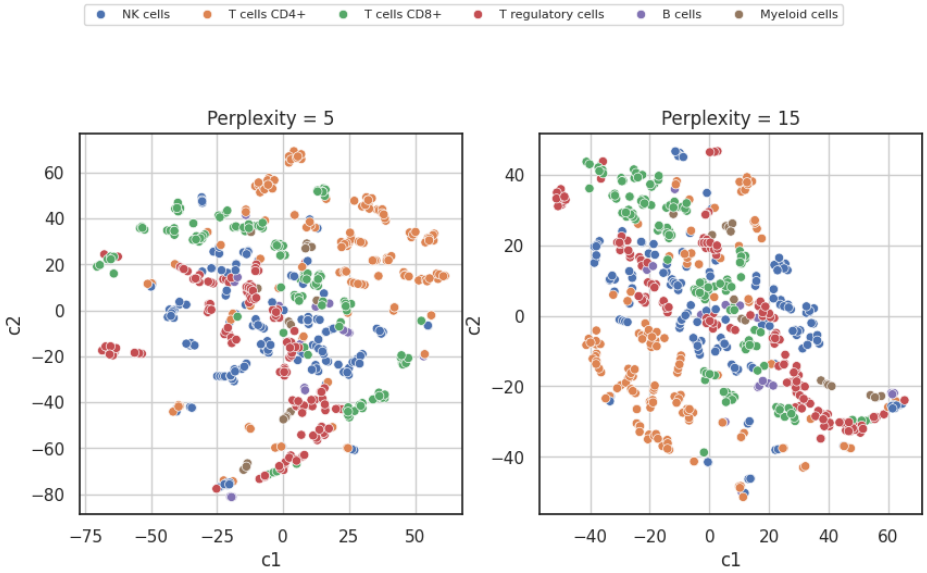

Рис. 2.1 Визуализация наблюдений экспрессии в пространстве размерности 2 с perplexity = 5, 10

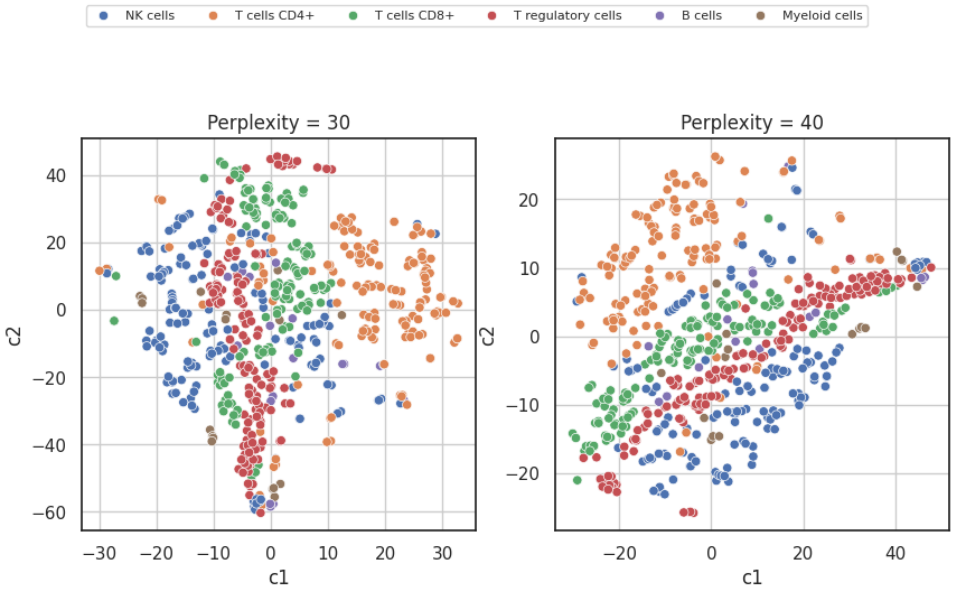

Рис. 2.2 Визуализация наблюдений экспрессии в пространстве размерности 2 с perplexity = 30, 40

Из представленных графиков можно наблюдать, что визуально выделяются кластеры вытянутой формы, которые тесно связаны с типом клеток; это более заметно при параметре перплексии равным 30 и 40.

2.2. Применение метода UMAP

Еще одним методом нелинейного снижения размерности является UMAP (Uniform Manifold Approximation and Projection). Алгоритм основан на подходах, полученных в результате анализа топологических данных и состоит из двух этапов: построение графа в больших размерностях и этапа оптимизации, иначе - поиск наиболее похожего графа в меньших размерностях. Алгоритм опирается на построение комплекса Чеха и строит нечеткий симплектический комплекс – это такое представление взвешенного графа, где веса ребер представляют вероятность того, что выбранные точки соединены. Вокруг каждой точки метод строит окружность – если окружности пересекаются, точки соединяются. Радиус такой окружности является довольно важным гиперпараметром: если окружности будут небольшими, то вершин в каждой из компонент связности графа будет немного, иначе образованные кластеры будут небольшими; в то же время при выборе большого радиуса образует слишком большие по набору наблюдений кластеры. Инструмент UMAP решает проблему выбора радиуса, выбирая его локально на основе расстояния до n-го ближайшего соседа. Далее алгоритм делает график «нечетким», уменьшая вероятность соединения по мере увеличения радиуса. При учете того, что каждая из точек должна быть соединена со своим ближайшим соседом метод гарантирует, что локальная и глобальная структура не противоречат друг другу. После построения многомерного графа UMAP оптимизирует компоновку аналога с низкой размерностью, чтобы он был как можно более похожим [15].

Одними из основных гиперпараметров алгоритма UMAP являются количество ближайших соседей и минимальное расстояние между точками в пространстве низкой размерности. Минимальное расстояние регулирует, насколько плотно алгоритм объединяет точки. Большие значения параметра позволит сохранить широкую топологическую структуру в данных, а выбор

низкие значений приводит к более плотной компоновке.

По сравнению с алгоритмом t-SNE UMAP часто лучше сохраняет структуру данных в пространстве низкой размерности, а также выполняет процедуру снижения размерности за меньшее время. Оба метода искажают многомерную форму, поэтому расстояния между наблюдениями в более низких измерениях, чем исходное, не поддаются прямой интерпретации, в отличии от линейных методов снижения размерности [15].

При решении задачи о снижении размерности с помощью метода UMAP были получены результаты работы алгоритмов с гиперпараметром n_neighbors (количество ближайших соседей), который равен 5, 50, 1000, 3000 соответственно. Результаты работы метода представлены на рис. 2.3 и рис 2.4.

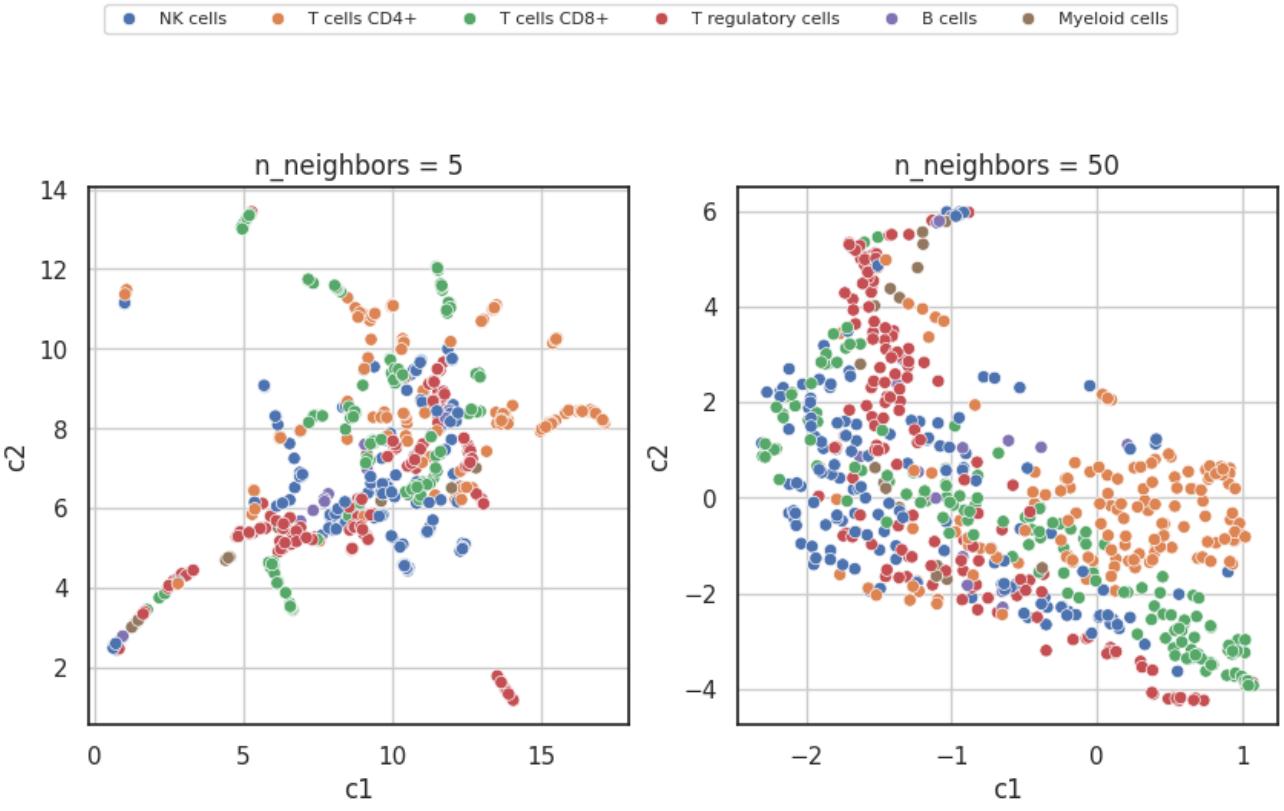

Рис. 2.3 Визуализация наблюдений экспрессии в пространстве размерности 2 с n_neighbors = 5, 50

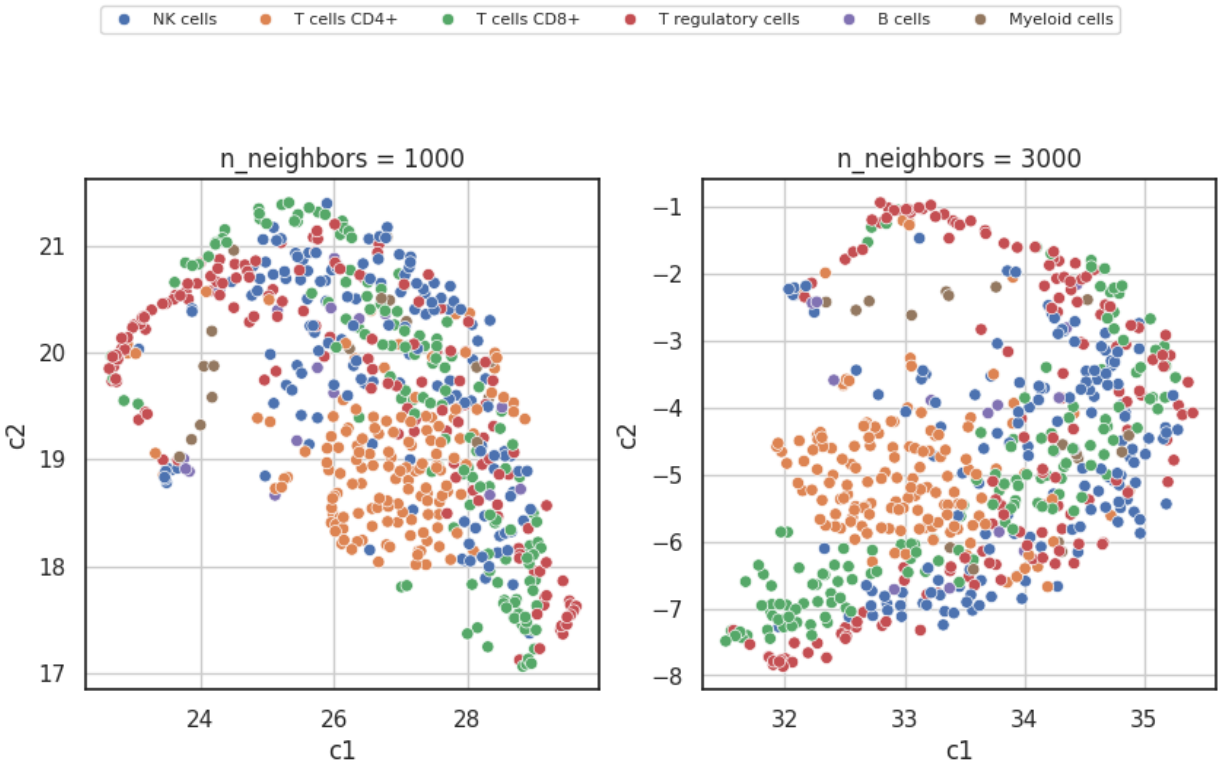

Рис. 2.4 Визуализация наблюдений экспрессии в пространстве размерности 2 с n_neighbors = 1000, 3000

Как и в случае применения метода t-SNE на графиках можно наблюдать кластеры, которые ассоциированы с типом клетки. Это значит, что различные лекарственные препараты могут вызывать похожую реакцию среди клеток периферической крови одного и того же типа.

2.3. Метод анализа главных компонент и LDA

В пункте 2.1 данной главы рассматривались нелинейные методы снижения размерности, которые хорошо подходят для визуализации данных с точки зрения сохранения пропорций расстояний между наблюдениями. Однако существуют алгоритмы преобразования данных, к которым не предъявляется требование сохранения близости похожих и отдаленности разных наблюдений при снижении размерности. Одним из таких методов является метод главных компонент – метод линейного преобразования, который использует разложение данных по сингулярным значениям [4] для проецирования их в пространство меньшей размерности; новые признаки будут являться линейными комбинациями исходных:

$$y_j = u_{j1}x_1 + u_{j2}x_2 + \cdots u_{j2}x_2, \quad (7)$$

Дисперсия выборки, которая посчитана относительно новых признаков является мерой того, как много информации удалось сохранить после понижения размерности, поэтому дисперсия должна быть максимальной. Для снижения размерности требуется найти компоненты u_i, на которые проецируются исходные данные. Будем искать такие u_i, что:

1. Компоненты ортогональны, иначе $(u_i, u_j) = 0$ при i≠j.
2. $||u_i|| = 1$.
3. Дисперсия проекции выборки на них максимальна: $D(Xu_i) \rightarrow max_{u_i}$, i = 1,..,.d, где Xu_i – проекция выборки X на компоненту u_i.

Для проведения дальнейших преобразований необходимо центрировать данные – вычесть из каждого признака его среднее, чтобы новое среднее было равно 0. Требование на максимизацию дисперсии проекции выборки на подпространство $\{u_1, u_2, \ldots u_d\}$ выглядит следующим образом:

$$\sum_{i=1}^{d} \left|\left|Xu_i\right|\right|^2 \rightarrow max_u, \quad (8)$$

Исходя из требований выше, можно получить, что первая компонента u_1 – собственный вектор матрицы X^TX с максимальным собственным значением, u_2 – собственный вектор матрицы X^TX со следующим по величине собственным значением и так далее. В таком случае доля объясненной дисперсии первыми k-компонентами равна

$$\delta_k = \frac{\lambda_1 + \lambda_2 + .. \lambda_k}{\lambda_1 + \lambda_2 + .. \lambda_n}, \quad (9)$$

где $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_n \geq 0$ – собственные числа матрицы X^TX в упорядоченном виде; 1-δ_k – доля необъясненной дисперсии.

Применим PCA с количеством компонент, равным 2 к данным об экспрессии генов и изобразим получившиеся компоненты на плоскости; результат представлен на рис.2.5.

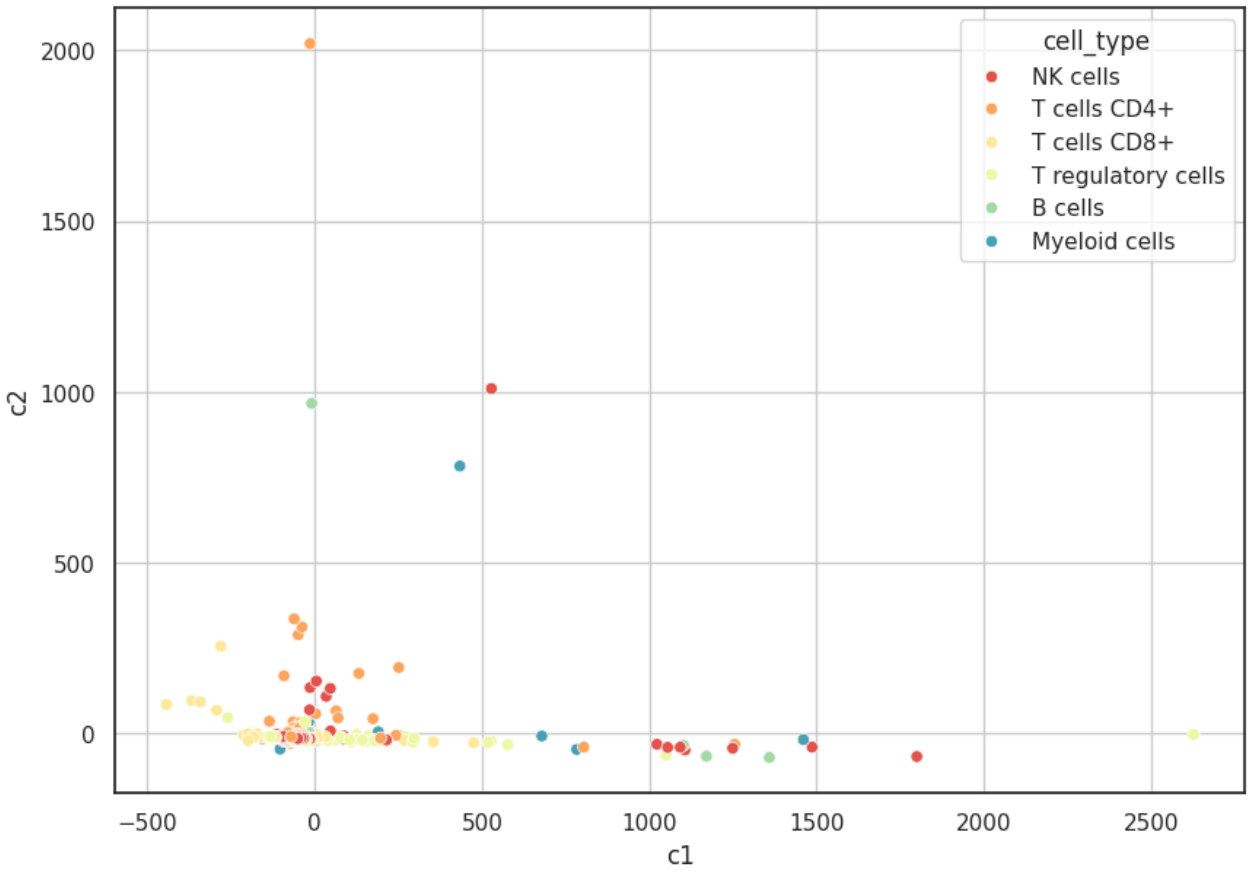

Рис. 2.5 Визуализация компонент, полученных в результате PCA

В отличии от нелинейных методов снижение размерностей, на графике, отражающем компоненты, полученные с помощью PCA не видно кластеров, которые ассоциированы с типом клеток, что может говорить о том, что PCA не стоит применять в задачах визуализации многомерных данных.

Наблюдения экспрессии можно разделить по классам, а именно по типам клеток, на которое действовало определенное соединение. Факт того, что в одинаковых типах клеток, подтверждает визуализация компонент, которые были получены методами нелинейного снижения размерности. Алгоритмы t-SNE и UMAP не учитывали информацию о наличии определенных классов внутри выборки и являлись алгоритмами обучения без учителя, как и PCA. Одним из линейных методов снижения размерности, задача которого минимизировать внутриклассовый разброс точек и максимизировать межклассовое расстояние в пространстве признаков является линейный дискриминантный анализ (LDA) [8]. В отличии от предыдущих методов LDA является методом обучения с учителем

– на вход методу подается не только выборка, но и вектор классов, где каждое значение соответствует определенному наблюдению. Метод подбирает такую проекцию, чтобы наилучшим образом разделить точки нескольких классов.

Рассмотрим случай с двумя классами C_1, C_2. Пусть N_1, N_2 – количество объектов в этих классах. Вычислим центры классов как

$$m_1 = \frac{1}{N_1} \sum_{x \in C_1} x\,, \tag{10}$$

$$m_2 = \frac{1}{N_2} \sum_{x \in C_2} x\,, \tag{11}$$

Проекцию данных x на вектор w можно записать как $w^T x$. Вычислив, куда попадут центры классы при проекции, получим:

$$\mu_1 = w^T m_1,, \tag{12}$$

$$\mu_2 = w^T m_2. \tag{13}$$

.

Определим дисперсию внутри каждого из проецированных классов:

$$s_1 = \sum_{x \in C_1} (w^T x - \mu_1)^2\,, \tag{14}$$

$$s_2 = \sum_{x \in C_2} (w^T x - \mu_2)^2\,. \tag{15}$$

В методе LDA в случае хорошей проекции должно выполняться:

- $(\mu_1 - \mu_2)^2$ максимально; классы должны быть максимально разделены
- ${s_1}^2 + {s_2}^2$ минимально; объекты каждого класса должны быть компактно расположены

Итоговым функционалом для обучения является критерий LDA Фишера:

$$J = \frac{(\mu_1 - \mu_2)^2}{{s_1}^2 + {s_2}^2} \to max_w, \tag{16}$$

Результат работы метода LDA на данных об экспрессии в каждой из клеток представлен на рис. 2.6.

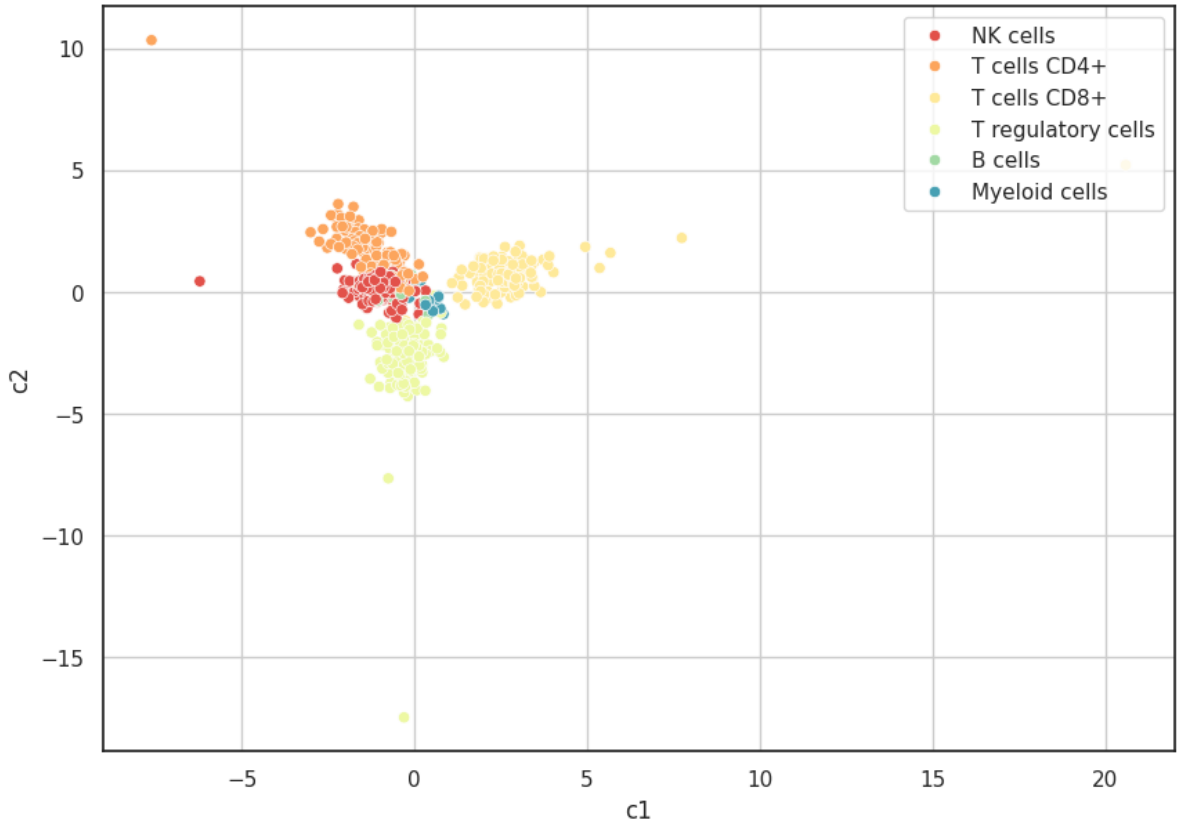

Рис. 2.6 Визуализация компонент, полученных в результате LDA

Из представленного графика можно сделать вывод, что метод хорошо справляется с задачей разделения классов: точки в двумерном пространстве образуют кластеры, которые ассоциированы с типом клеток периферической крови человека.

ГЛАВА 3. АНАЛИЗ ДАННЫХ СТАТИСТИЧЕСКИМИ ТЕСТАМИ

3.1 Оценка различий между группами клеток

Результаты визуализация данных об экспрессии генов показывают, что наблюдения разбиваются на группы, которые ассоциированы с типом клеток. Чтобы определить, действительно ли свойства определенной клетки влияют на конечную экспрессию, можно использовать определенные статистические критерии. Гипотезы, которые тестируют эти критерии, а также область применения отличаются: для применения некоторых из них к данным предъявляется требование нормального распределения данных; также не все из критериев проверяют равенство средних между группами. В данной задаче для сравнения нескольких групп можно было бы применить однофакторный дисперсионный анализ (ANOVA) – метод статистического анализа данных, который используется для определения статистически значимых различий между двумя и более группами по одной независимой переменной. Критерий проверяет нулевую гипотезу о том, что среднее значение зависимой переменной одинаково во всех группах. Однако метод требует, чтобы данные имели нормальное распределение [7] – в главе 1 для каждой из групп был проведен тест Шапиро-Уилка, результаты которого говорят о том, что каждая из выборок была взята не из нормального распределения. Поэтому стоит рассмотреть использование непараметрических критериев, которые не требуют нормального распределения данных в каждой из групп. Такие статистические тесты основаны на рангах – иначе, вместо выборочных значений используется их ранги (номера элементов в упорядоченной по возрастанию выборке). Одним из таких критериев является критерий Манна-Уитни. Для использования данного метода нужно, что выполняется о независимости наблюдений обеих групп друг от друга. Нулевая гипотеза критерия заключается в том, что распределение средней экспрессии одинаково для двух групп, альтернативная - распределение средней экспрессии различно для двух групп [1].

Для проведения тестирование гипотез, необходимо выполнить следующие шаги:

1. Составить единый ранжированный ряд из обеих сравниваемых выборок, расставив элементы по степени возрастания признака и приписав наименьшему значению наименьший номер – ранг.
2. Посчитать сумму рангов для первой и второй выборок.
3. Определить наибольшую из ранговых сумм T.
4. Вычислить эмпирическое значение U критерия:

$$U = n_x * n_y + \frac{n(n+1)}{2} - T, \quad (17)$$

где n_x, n_y – объемы выборки, n – объем выборки, имеющей наибольшую ранговую сумму.

5. Для заданного уровня значимости (p-value = 0.05) сравниваем U с критическим значением. Если значение статистики меньше критического, можем отвергнуть нулевую гипотезу о равенстве распределений в группах.

В данных средней экспрессии присутствуют выбросы, что иллюстрируют скрипичные диаграммы и графики boxplot. В отличии от критериев, где важную роль играет распределение и отсутствие шумов, критерий Манна-Уитни способен находить статистически значимый эффект на данных такого рода.

В эксперименте проводится сравнение между шести группами; тест Манна-Уитни способен проводить попарные сравнения, всего сравнений будет 15. Из-за того, что данные не распределены нормально, использование ANOVA, который позволяет сравнивать сразу несколько групп, будет некорректным. Альтернативой дисперсионному анализу выступает критерий Краскела-Уоллиса, который, как и тест Манна-Уитни, является непараметрическим и основан на рангах. Для проведения расчетов необходимо вычислить статистику H. Проверяется нулевая о том, что медиана средней экспрессии одинакова для каждой из групп и альтернативная гипотеза, которая гласит, что медиана средней экспрессии, по крайней мере, отличается в одной из выборок.

Статистика критерия H вычисляется по формуле [5, с. 586]:

$$H = \frac{12}{n(n+1)} \sum \frac{R_i^2}{n_i} - 3(n+1), \quad (18)$$

где n – общее количество наблюдений во всех выборках, R_i – сумма рангов для i группы.

Как и в случае критерия Манна-Уитни, статистика сравнивается с критическим значением для заданного количества степеней свободы (n-1) и p-value.

3.2 Непараметрические критерии для сравнения групп клеток

С помощью средств программного пакета scipy для языка программирования Python проведем тест Краскела-Уоллиса и получим значение H-статистики и p-value:

- H = 47.17
- p-value = $5.23*10^{-9}$

Приняв уровень статистической значимости равен 0.05, можно отклонить нулевую гипотезу о том, что во всех группах одинаковое значение медианы. Иначе, есть как минимум одна группа, медиана средней экспрессии генов которой отличается от остальных. Критерий Краскела-Уоллиса не отвечает на вопрос, какие именно из групп отличаются. Сравнить каждую выборку с каждой можно с помощью критерия Манна-Уитни, всего будет 15 попарных сравнений. Результаты сравнения, которые представляют из себя U-статистику и p-value, приведены в таблице 3.1.

	U-статистика	p-value
NK-клетки, Т-клетки (CD4+)	13172	$4.93*10^{-4}$
NK-клетки, Т-клетки (CD8+)	13259	$4.25*10^{-5}$

NK-клетки, Т-клетки (регуляторные)	9396	0.08
NK-клетки, миелоидные клетки	822	0.02
NK-клетки, B-клетки	1033	0.26
Т-клетки (CD4+), Т-клетки (CD8+)	11706	0.06
Т-клетки (CD4+), Т-клетки (регуляторные)	7945	$1.70*10^{-4}$
Т-клетки (CD4+), миелоидные клетки	571	$2.78*10^{-4}$
Т-клетки (CD4+), B-клетки	679	$2.30*10^{-3}$
Т-клетки (CD8+), Т-клетки (регуляторные)	7012	$2.08*10^{-6}$
Т-клетки (CD8+), миелоидные клетки	540	$2.03*10^{-4}$
Т-клетки (CD8+), B-клетки	726	$7.40*10^{-4}$
Т-клетки (регуляторные), миелоидные клетки	949	0.11
Т-клетки (регуляторные), B-клетки	1183	0.75
B-клетки, миелоидные клетки	164	0.51

Таблица 3.1. Результаты теста Манна-Уитни

В 6 из 15 случаев при уровне значимости 0.05 мы не можем отвергнуть нулевую гипотезу о том, что в сравниваемых группах распределение средней экспрессии одинаково для обеих групп – клетки могут иметь одинаковое происхождение.

ГЛАВА 4. ПРИМЕНЕНИЕ АЛГОРИТМОВ РЕГРЕСИИ

4.1 Модель LightGBM

Разведочный анализ данных и применение статистических методов помогли больше узнать об исследуемых данных, выявить ассоциации и оценить влияние каждого из факторов. В то же время для прогнозирования экспрессии применяются методы машинного обучения, которые, в данном случае, нацелены на решение задачи регрессии. Постановка задачи регрессии имеет следующий вид:

- Пусть X - множество объектов, Y – множество ответов
- y: X→Y – неизвестная зависимость
- $\{x_1, x_2, \dots x_l\} \subset X$ – обучающая выборка, $y_i = y(x_i)$, где i = 1..l – известные ответы.
- Требуется найти a: X→Y – алгоритм (решающую функцию), приближающую y на всем множестве X.

Для оценки результатов работы модели используются различные метрики, среди которых RMSE, MSE, R^2. Формулы для вычисления:

$$\mathrm{MSE} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\mathrm{y_i} - \mathrm{f(x_i)}\right)^2, \tag{19}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\mathrm{y_i} - \mathrm{f(x_i)}\right)^2}, \tag{20}$$

$$\mathrm{R^2} = 1 - \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\mathrm{y_i} - \mathrm{f(x_i)}\right)^2}{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y_i} - \bar{\mathrm{y}})^2}, \tag{21}$$

где y_i – известные ответы, $f(x_i)$ – ответы, полученные в результате работы модели, $\bar{y}$- среднее арифметическое отметок y.

В данной задаче необходимо спрогнозировать не одну, а 18211 переменных, поэтому решается задача многовыходной регрессии. Метрикой является MRRMSE – среднеквадратичная ошибка по строкам:

$$\mathrm{MRRMSE} = \frac{1}{\mathrm{R}}\sum_{\mathrm{i}=1}^{\mathrm{R}}\sqrt{\frac{1}{\mathrm{n}}\sum_{\mathrm{j}=1}^{\mathrm{n}}\left(\mathrm{y_{ij}} - \mathrm{f}\left(\mathrm{x_{ij}}\right)\right)^2}, \quad (22)$$

где R – количество столбцов, n – количество строк, $\mathrm{y_{ij}}$, $\mathrm{f}\left(\mathrm{x_{ij}}\right)$ – фактические и прогнозные значения экспрессии соответственно.

Одним из сравниваемых алгоритмов регрессии в данной задаче является модель LightGBM, основанная на градиентном бустинге. Базовой моделью алгоритма является решающее дерево $b_i(x)$. Каждое дерево $b_i(x)$ в композиции строится так, чтобы уменьшить ошибку на предыдущем шаге. Итоговым алгоритмом будет композиция из N деревьев, где N – гиперпараметр [12]:

$$a_i(x) = \sum_{i=1}^{N} b_i(x)\,. \quad (23)$$

Функция потерь для дерева решений с индексом j имеет следующий вид:

$$L(y,x) = \sum_{i=1}^{n}\left((s_j)_i - b_j(x_i)\right)^2 \to min\,, \quad (24)$$

где i – номер объекта обучающей выборки, s_j – вектор ошибок, полученный на предыдущем шаге:

$$s_j = \mathrm{y} - \sum_{i=1}^{j} b_i(x)\,. \quad (25)$$

С другой стороны, чтобы функция потерь гарантировано уменьшалась на каждом шаге, вектор ошибки стоит вычислять как градиент функции потерь с обратным знаком:

$$s_i = -\frac{\partial L}{\partial a_{N-1}(x)}. \quad (26)$$

Таким образом, при добавлении предсказания дерева решений в итоговый ответ, функция потерь уменьшится засчет движения по антиградиенту функции потерь.

Для улучшения результатов прогнозирования, а именно для уменьшения метрики MRRMSE важно подобрать гиперпараметры модели LightGBM. Так как

базовым алгоритмом является дерево решений, алгоритм подбирает следующие гиперпараметры:

- Глубина и количество деревьев.
- Вероятность разделения выборки .
- Минимальное количество данных в листе.

Самому методу можно установить количество параллельных потоков, используемых для обучения и параметр learning_rate, который используется для регулирования скорости обучения. Для каждого из алгоритмов LightGBM подбор констант осуществлялся с помощью метода Optuna, который основан на реализации идеи байесовского подбора гиперпараметров. До начала оптимизации настраивается целевая функция, которая на вход получает специальный объект для определения пространства поиска параметров. Затем Optuna реализует метод Tree-Structured Parzen Estimator, который для оптимизации реализует следующие шаги:

- Подбор группы значений определенного гиперпараметра, на которых известно качество модели с помощью случайного поиска
- Разбиение данных на две группы – объекты, для которых модель показала лучшее качество и все остальные объекты; доля лучших наблюдений также является гиперпараметром. Для групп строятся оценки распределения l(x) и g(x) соответственно.
- Для каждого из нескольких значений кандидатов из l(x) рассчитывается ожидаемое улучшение (Expected Improvement):

$$\mathrm{E}I = \frac{l(x)}{g(x)} \tag{27}$$

- После выбора значения с максимальным ожидаемым улучшением обучается модель с гиперпараметром, который принимает это значение.

Алгоритм оптимального поиска гиперпараметров применяется для каждой из 18211 моделей отдельно – в качестве минимизируемой метрики регрессии используется RMSE, потому что информация для каждого из отдельных генов представляет из себя числовой вектор. Итоговый датасет разделяется на

обучающую и валидационную выборку в соотношении 3:1; для оптимизации целевой функции используется 10 итераций.

С точки зрения вычислительных возможностей обучение алгоритма регрессии является трудоемкой задачей – после извлечения новых данных из свойств лекарственных препаратов обучающая выборка имеет 146 признаков; в то же время решается задача обучения 18211 таких алгоритмов – в соответствии с количеством генов. Одним из вариантов ускорения процесса обучения является снижение размерности исходных данных с помощью метода главных компонент. В то же время метод целесообразно применять, если при небольшом количестве получившихся компонент удается сохранить большую долю объясненной дисперсии. На рис.4.1 изображена зависимость количества компонент и суммарной доли объясненной дисперсии – из графика можно сделать вывод, что без потери суммарной доли объясненной дисперсии можно убрать 15 компонент, что составляет около 10% от общего числа признаков, что не сильно уменьшит время вычислений.

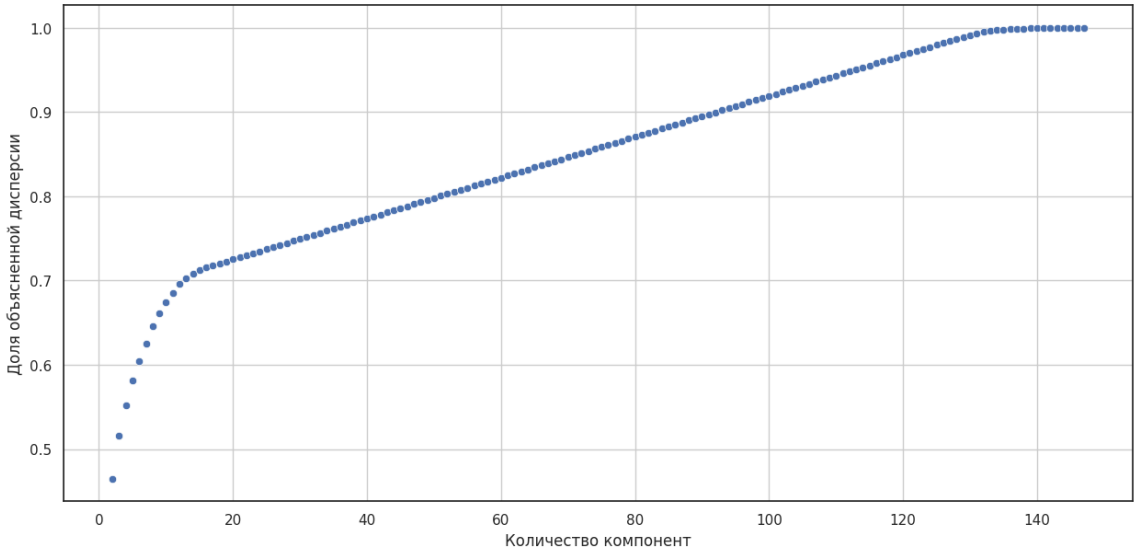

Рис. 4.1 Зависимость доли объясненной дисперсии от количества компонент

Для решения задачи регрессии использовался алгоритм LightGBM в нескольких вариациях: без использования Optuna, с использованием Optuna для предсказания информации по отдельным генам – алгоритм подбора гиперпараметров применялся в 1%, 5% и 100% случаев. Результаты применения

алгоритма представлены в таблице 4.1.

Модификация LightGBM	MRRMSE
Без применения Optuna	0.706
С применением Optuna в 1% случаев	0.706
С применением Optuna в 5% случаев	0.705
С применением Optuna в 100% случаев	0.670

Таблица. 4.1 Результаты работы алгоритма LightGBM

Обучение алгоритма, где гиперпараметры подбирались для всех алгоритмов, длилось около 6 часов - несмотря на то, что метод Optuna требователен к вычислительным ресурсам, его применение уменьшает метрику MRRMSE и делает итоговый алгоритм лучше для решения подобных задач регрессии.

4.2. Модель Py-Boost

Одной из моделей, применяемой в данной задаче, является py-boost – метод, как и LightGBM, основан на идее градиентного бустинга. Для обучения алгоритм использует только графический процессор и пользуется библиотеками языка программирования Python для работы с GPU, такие как CuPy и Numba [9], что отличает данный метод от остальных методов, где базовой моделью является решающее дерево. Применение py-boost требует больших вычислительных ресурсов; оценим результаты регрессии в рамках одного гена. На рис.4.2 представлена кривая обучения – график зависимости минимизируемой метрики от размера обучающих данных в зависимости от количества решающих деревьев в модели. В качестве метрики выступает RMSE. Из представленной зависимости можно сделать вывод о том, что количество решающих деревьев можно оставить заданными по умолчанию, потому что метрика при возрастании количества ведет себя нестабильно; изначально – 100 базовых моделей.

Рис. 4.2 Кривая обучения py-boost для одного гена

При использовании модели py-boost с количеством деревьев решений, равным 100, результат регрессии, выраженный в метрике MRRMSE, равен 0.722, время обучения составляет около 10 часов – итог получился хуже, чем у предыдущего алгоритма LightGBM.

4.3. Ансамблирование алгоритмов регрессии

Алгоритмы регрессии, использующие разные подходы для решения задач машинного обучения, находят разные закономерности в данных. В связи с этим, одной из идей улучшения метрик является использование нескольких моделей регрессии, а затем усреднение результатов. Общее время работы алгоритмов также не должно быть высоким, поэтому предпочтительно использование простых моделей. В качестве таких алгоритмов выступают метод к-ближайших соседей и модификация метода опорных векторов для решения задач регрессии. Метод опорных векторов использует линейное ядро [3], применяемое к преобразованию признаков и представимое в следующем виде:

$$K(x_1, x_2) = (x_1, x_2), \tag{28}$$

(x_1, x_2) – скалярное произведение векторов x_1, x_2.

В качестве метода ближайших соседей используется модель, также модифицированная под решение задач регрессии с разным количеством соседей. Выбрать коэффициенты, с которыми суммируются результаты разных моделей, можно либо эмпирически, либо с помощью линейной регрессии. Порядок проведения расчетов выглядит следующим образом:

- Определяется один экземпляр класса метода опорных векторов и несколько экземпляров классов метода к-ближайших соседей с количеством соседей 3,5,7…19.
- Исходный датасет на тренировочную и валидационную часть.
- После этого алгоритмы обучаются на тренировочной части и делается прогноз на валидационной части.
- Получившиеся прогнозы представляют из себя вектора – из них составляется новая обучающая выборка, где каждое из получившихся предсказаний будет являться признаком; в то же время вектор валидационных (истинных) значений экспрессии генов будет являться зависимой переменной
- Задача получившейся обучающей выборки – обучить модель линейной регрессии:

$$\text{y} = intercept + \sum_{i=1}^{N} \beta_i x_i \,, \tag{29}$$

где y – зависимая переменная, $intercept$ – свободный член в уравнении линейной регрессии, x_i – признаки, которые получились в результате обучения алгоритмов на валидационной выборке, β_i – коэффициенты при признаках, N – количество различных алгоритмов регрессии.

- Далее обучаем базовые алгоритмы на всей выборке, получаем предсказания нескольких алгоритмов на тестовом датасете.
- Итоговым ответом будет являться:

$$y_{predicted} = intercept * e + \sum_{i=1}^{N} \beta_i z_i \,, \tag{30}$$

где e – вектор из единиц с размерностью R^n, n – количество наблюдений в исходных данных, z_i – векторы, которые является результатом прогнозирования базовых алгоритмов регрессии, $y_{predicted}$ – вектор итоговых ответов.

У решения этой задачи есть и другой подход: представить итоговый ответ как композицию предсказаний разных моделей с весами, которые в сумме равны единице:

$$y_{predicted} = \sum_{i=1}^{N} \alpha_i z_i \,, \tag{31}$$

$$\sum_{i=1}^{N} \alpha_i = 1 \,, \tag{32}$$

где α_i – коэффициенты, подобранные эмпирически, z_i – векторы, которые является результатом прогнозирования базовых алгоритмов регрессии. Сравнение двух подходов приведено в таблице 4.2.

	MRRMSE
Коэффициенты подобраны с помощью линейной регрессии	0.715
Эмпирический подбор коэффициентов	0.602

Таблица. 4.2 Результаты ансамблирования нескольких алгоритмов регрессии

Наибольшим из коэффициентов при использовании второго подхода выбирался коэффициент для ядрового метода опорных векторов; он варьировался от 0.25 до 0.4. Модифицированным алгоритмам кластеризации присваивался весовой коэффициент, который менялся от 0.09 до 0.15. Ансамблирование алгоритмов регрессии дает наилучший результат из рассматриваемых алгоритмов, потому что является усреднением результатов работы методов, которые выявляют в данных разные закономерности.

ЗАКЛЮЧЕНИЕ

В ходе данной работы были изучены данные об экспрессии генов в клетках периферической крови человека, а также математический аспект биологического эксперимента, в результате которого были получены данные. Имеющийся набор данных был исследован на наличие аномалий и выбросов, также были построены графики, объясняющую структуру датасета; для улучшения алгоритмов регрессии были спроектированы новые признаки на основе свойств различных лекарственных соединений.

Далее, для выявления ассоциаций между экспрессией генов и признаками в данных были применены алгоритмы снижения размерности, основанные как на нелинейных, так и на линейных методах, в ходе чего была выявлена взаимосвязь между средней экспрессией и типом клетки, в которой была зафиксирована средняя экспрессия. Сделан вывод о том, что для визуализации многомерных данных больше подходят методы, сохраняющие пропорции расстояний между объектами или алгоритмы снижения размерности, основанные на обучении с учителем.

Также проведен ряд непараметрических статистических тестов для оценки различий между группами клеток. Непараметрические методы были выбраны потому, что данные о средней экспрессии не распределены нормальным образом; в данных также присутствуют выбросы. Результаты применения статистических критериев показали, что группы действительно отличаются между собой – это может быть полезно при настройке методов машинного обучения для улучшения результатов предсказания экспрессии на новом наборе данных.

Проведено сравнение алгоритмов регрессии, основанные на разных подходах: решающие деревья, градиентный бустинг, ансамблирование нескольких различных методов. Наилучшим решением оказалась композиция результатов регрессии ядрового метода опорных векторов и метода k-ближайших соседей, которые были модифицированы для решения регрессионных задач.

Направления продолжения исследования могут быть следующими:

- Применение методов глубокого обучения для предсказания экспрессии на новом наборе данных.
- Улучшение модели путем добавление количественных данных о клетках периферической крови человека.
- Проверка на устойчивость к шуму во входных данных и объяснение шума с клинической точки зрения.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Большев Л. Н., Смирнов Н. В., Таблицы математической статистики // 2 изд., М. : Наука, 1983. – 416 с.
2. Кобзарь А. И. Прикладная математическая статистика // М.: Физматлит, 2006. — 816 с.
3. Плас Дж. Вандер Python для сложных задач: наука о данных и машинное обучение // ООО Издательство «Питер», 2022. — 576 с.
4. Geron A., Hands on Machine Learning with Scikit Learn, Keras and Tensorflow Concepts Tools and Techniques to Build Intelligent Systems //O'Reilly Media, 2019 — P. 510.
5. Kruskal W. H. and Wallis W. A. Use of ranks in one-criterion variance analysis. // Journal of the American Statistical Association. — 1952, т. 47 №260, с. 583–621.
6. Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, Gordon K. Smyth. Limma powers differential expression analyses for RNA-sequencing and microarray studies // Nucleic Acids Research, 2015, т. 43, № 7, с.47.
7. Дисперсионный анализ (ANOVA) [Электронный ресурс]. Режим доступа: https://habr.com/ru/companies/otus/articles/734258/ (дата обращения: 11.05.2024).
8. Линейный дискриминантный анализ (LDA) [Электронный ресурс]. Режим доступа: https://habr.com/ru/articles/802511/ (дата обращения: 08.05.2024).
9. Обучение моделей методом градиентного бустинга на GPU [Электронный ресурс]. Режим доступа: https://developers.sber.ru/portal/products/py-boost (дата обращения: 01.05.2024).
10. Одноклеточное секвенирование: разделяй, изучай и властвуй [Электронный ресурс]. Режим доступа: https://biomolecula.ru/articles/odnokletochnoe-sekvenirovanie-razdeliai-izuchai-i-vlastvui (дата обращения: 11.05.2024).
11. Стохастическое вложение соседей с t- распределением [Электронный

ресурс]. Режим доступа: https://neerc.ifmo.ru/wiki/index.php?title=Стохастическое_вложение_соседей_с_t-распределением (дата обращения: 08.05.2024) .

12. Учебник по машинному обучению ШАД [Электронный ресурс]. Режим доступа: https://education.yandex.ru/handbook/ml (дата обращения: 01.04.2024).
13. Connectivity Map (CMAP) [Электронный ресурс]. Режим доступа: https://www.broadinstitute.org/connectivity-map-cmap (дата обращения: 11.05.2024).
14. Differential gene expression analysis [Электронный ресурс]. Режим доступа: https://www.sc-best-practices.org/conditions/differential_gene_expression.html (дата обращения: 01.05.2024).
15. Understanding UMAP [Электронный ресурс]. Режим доступа: https://pair-code.github.io/understanding-umap/ (дата обращения: 08.05.2024).