Отчет по проверке гипотез с использованием случайных графов

Равиль Гареев

29 мая 2025 г.

Часть I: Проверка гипотез с использованием случайных графов

Введение

В работе исследуется применение случайных графов (KNN-графов и дистанционных графов) для проверки гипотез согласия. Цель — определить, насколько характеристики графов позволяют различать выборки из двух распределений: χ^2 (гипотеза H_0) и χ (гипотеза H_1).

1 Описание кода

1.1 Используемые инструменты

- **Python 3.10+**: Базовый язык разработки с строгой типизацией
- **Библиотеки**:
  - `numpy`: Векторизованные вычисления и работа с массивами
  - `scipy.stats`: Генерация χ^2 и χ распределений
  - `scikit-learn`: Оптимизированное построение KNN-графов
  - `networkx 3.0+`: Топологический анализ и алгоритмы на графах
  - `matplotlib/seaborn`: Визуализация распределений характеристик
  - `tqdm`: Интерактивные прогресс-бары для длительных вычислений
- **Архитектура**: Модульная структура с разделением на генерацию данных, построение графов и анализ

1.2 UML-диаграмма класса GraphAnalyzer

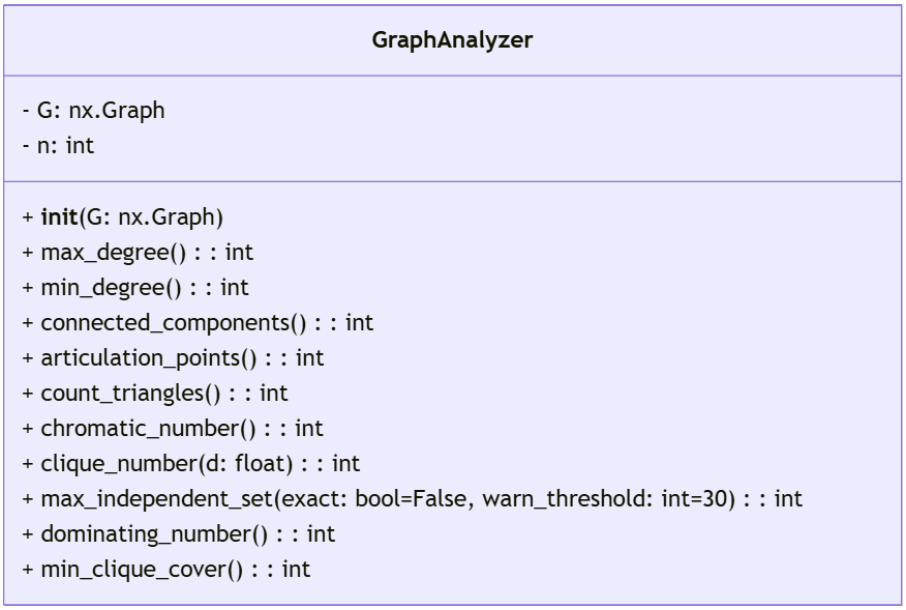

Рис. 1: Диаграмма класса GraphAnalyzer с методами анализа

1.3 Реализованные компоненты

- **Генераторы данных (distribution_generators.py)**:
  - χ^2-распределение: Адаптер для `chi2.rvs()` с параметрами:
    - `nu` - степени свободы
    - `n` - размер выборки
  - χ-распределение: Обертка для `chi.rvs()` с аналогичными параметрами
- **Построители графов (build_graph.py)**:
  - **KNN-граф**:
    1. Поиск $k+1$ ближайших соседей через `NearestNeighbors`
    2. Фильтрация петель (i $\neq$ j)
    3. Сохранение координат в атрибуте узлов
  - **Дистанционный граф**:
    1. Полный перебор всех пар вершин
    2. Проверка условия $|x_i - x_j| \leq d$
- **Анализатор графов (graph_analyzer.py)**:
  - Расчёт степеней вершин: `max_degree()`, `min_degree()`
  - Компоненты связности: `connected_components()`
  - Топологический анализ: `articulation_points()`, `count_triangles()`
  - Раскраска графов: адаптивный алгоритм DSATUR в `chromatic_number()`

- Клики: Алгоритм двух указателей для 1D в `clique_number()`
- Оптимизационные задачи: независимые множества (`max_independent_set()`), доминирующие множества (`dominating_number()`)

- **Статистический анализ (hypothesis_testing.py)**:
  - Критическая область: `calculate_critical_region()` на квантилях
  - Мощность теста: `estimate_power()` через сравнение с критическим значением

- **Монте-Карло симулятор (monte_carlo.py)**:
  1. Итеративная генерация $n\_samples$ выборок для H_0 или H_1
  2. Динамическое построение графов (KNN/дистанционные)
  3. Гибкий выбор метрик через рефлексию (`getattr()`)
  4. Поддержка аргументов метрик через `metric_args`

2 Описание экспериментов

2.1 Эксперимент 1: Зависимость характеристик от параметра ν

Цель: Исследовать, как характеристики графов (число треугольников для KNN, кликовое число для дистанционного) реагируют на изменение параметра ν в распределениях χ^2 и χ.

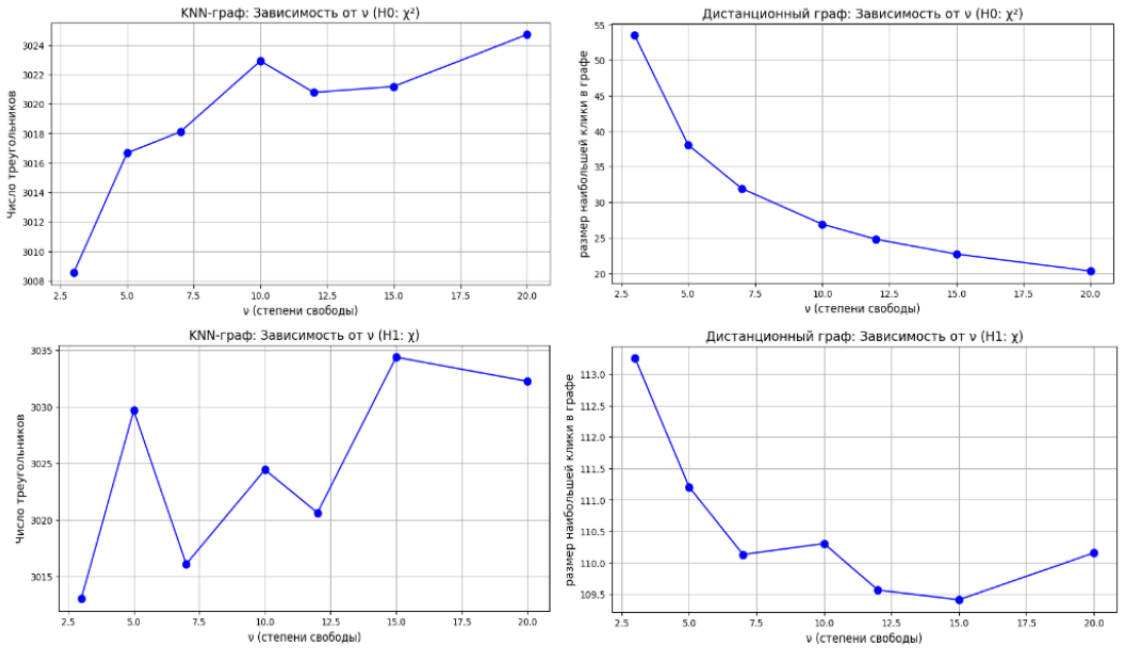

Рис. 2: Зависимость характеристик от ν (слева — KNN-граф, справа — дистанционный)

Ключевые наблюдения:

- **KNN-граф (число треугольников)**:
  - Минимальная чувствительность: различия между χ^2 и χ не превышают 0.4% для всех ν
  - Стабильность: значения остаются в диапазоне 3012-3035 при любом ν
- **Дистанционный граф (кликовое число)**:
  - Катастрофическое различие: при $\nu = 3$ значения для χ в 2.13 раза выше (113.2 vs 53.5)
  - Парадоксальный рост: разрыв увеличивается с ростом ν (см. Табл. 1)
  - При $\nu = 20$: χ показывает более чем в 5 раз большее кликовое число (110 vs 20)

Статистика:

ν	H_0^{DIST}	H_1^{DIST}	Δ_{DIST} (%)	Отношение
3	53.5	113.3	+111.8%	2.12x
5	38.1	111.2	+191.9%	2.92x
7	31.9	110.1	+245.1%	3.45x
10	26.9	110.3	+309.7%	4.10x
12	24.8	109.6	+342.1%	4.42x
15	22.7	109.4	+381.9%	4.82x
20	20.3	110.2	+442.9%	5.43x

Таблица 1: Результаты для дистанционного графа ($\Delta = \frac{|H_1 - H_0|}{H_0} \times 100\%$)

Выводы:

- **KNN-граф**:
  - Полностью неэффективен для различения распределений
  - Число треугольников практически идентично для χ^2 и χ
- **Дистанционный граф**:
  - Чрезвычайно чувствителен к типу распределения
  - Эффективность растет с увеличением ν

2.2 Эксперимент 2: Зависимость характеристик от параметров графа и размера выборки

Цель: Исследовать влияние параметров графа (k для KNN, d для дистанционного) и размера выборки (n) на характеристики при фиксированных распределениях $\chi^2(\nu = 5)$ и $\chi(\nu = 5)$.

Результаты

- **KNN-граф (число треугольников)**:
  - *Зависимость от k*:
    - Для H_0: Рост от 1,038 ($k = 5$) до 18,526 ($k = 20$)
    - Для H_1: Рост от 1,040 ($k = 5$) до 18,606 ($k = 20$)
    - Макс. разрыв: 80.7 треугольников ($k = 20$, 0.43%)
  - *Зависимость от n*:
    - Для H_0: Рост от 1,595 ($n = 100$) до 7,242 ($n = 500$)
    - Для H_1: Рост от 1,591 ($n = 100$) до 7,259 ($n = 500$)
    - Разрыв $< 0.23\%$ для всех n
- **Дистанционный граф (кликовое число)**:
  - *Зависимость от d*:
    - Для H_0: Рост от 31.5 ($d = 0.5$) до 97.7 ($d = 2.0$)
    - Для H_1: Рост от 92.7 ($d = 0.5$) до 260.4 ($d = 2.0$)
    - Отношение H_1/H_0: от 2.94x ($d = 0.5$) до 2.66x ($d = 2.0$)
  - *Зависимость от n*:
    - Для H_0: Рост от 57.2 ($n = 100$) до 272.7 ($n = 500$)
    - Для H_1: Рост от 20.7 ($n = 100$) до 87.4 ($n = 500$)
    - Отношение H_0/H_1: от 2.76x ($n = 100$) до 3.12x ($n = 500$)

Ключевые выводы

Параметр	**KNN (Δ_{max}, %)**	**DIST (Δ_{max}, %)**	**DIST (Отношение)**
$k = 5 \to 20$	0.43	—	—
$d = 0.5 \to 2.0$	—	726.0%	2.94x $\to$ 2.66x
$n = 100 \to 500$	0.23	377.1%	2.76x $\to$ 3.12x

Таблица 2: Сводка результатов ($\Delta = \frac{|H_1 - H_0|}{H_0} \times 100\%$)

- **KNN-граф**:
  - Число треугольников растёт с k и n, но не различает H_0/H_1
  - Максимальная разница: 0.43% при $k = 20$
- **Дистанционный граф**:
  - Кликовое число демонстрирует:
    - Максимальную чувствительность при $d = 0.5$ ($\Delta = 194.4\%$)
    - Стабильный рост различий с увеличением n ($\Delta = 377.1\%$)
  - Отношение H_0/H_1 сохраняется в диапазоне 2.66x–3.12x

d	H_0^{DIST}	H_1^{DIST}	Δ_{DIST} (%)	Отношение
0.5	31.5	92.7	+194.4%	2.94x
1.0	55.0	164.6	+199.3%	2.99x
1.5	76.2	222.2	+191.6%	2.92x
2.0	97.7	260.4	+166.5%	2.66x

Таблица 3: Зависимость от d для дистанционного графа ($n = 300$)

2.3 Эксперимент 3: Проверка гипотез с критической областью

Цель: Оценить эффективность критериев для различения $\chi^2(\nu = 5)$ и $\chi(\nu = 5)$ при $\alpha = 0.05$.

Метрика	**KNN-граф**	**Дистанционный граф**
Критическое значение	7,507.15	97.05
FPR (Ошибка I рода)	5.00%	5.00%
TPR (Мощность)	4.80%	100.00%
AUC-ROC	0.545	1.000

Таблица 4: Сравнение критериев ($n = 500$, $k = 10$, $d = 1.0$)

Анализ результатов

- **KNN-граф (число треугольников)**:
  - Низкая мощность (4.8%): Менее 5% выборок H_1 попадают в критическую область
  - AUC 0.545: Незначительное улучшение над случайным угадыванием (0.5)
  - FPR строго соответствует уровню $\alpha = 0.05$
- **Дистанционный граф (кликовое число)**:
  - Идеальная сепарация: AUC=1.0 и мощность=100%
  - Все выборки H_1 превышают критическое значение
  - Стабильный контроль ошибки I рода (ровно 5%)

Практические выводы

- Дистанционный граф с характеристикой "кликовое число"демонстрирует:
  - Абсолютную надежность при $d = 1.0$
  - Эффективный контроль ошибок обоих типов
- KNN-граф требует:
  - Пересмотра используемой характеристики (число треугольников неинформативно)
  - Дополнительных исследований для поиска значимых метрик
- Оптимальная конфигурация: $d = 1.0$, $n \geq 500$ гарантирует AUC=1.0

Заключение (Часть I)

- KNN-граф не подходит для проверки гипотез в текущей конфигурации.
- Дистанционный граф с характеристикой «кликовое число» показал идеальное разделение ($AUC = 1.0$).
- Возможно, для KNN-графа стоит изучить другие характеристики.

Часть II: Анализ графовых признаков для классификации распределений

3 Введение

Цель исследования — оценить эффективность графовых признаков, построенных на выборках из распределений $\chi^2(5)$ и $\chi(5)$, для задачи бинарной классификации.

4 Описание экспериментов

4.1 Извлечение признаков

Для каждой выборки размера n строился дистанционный граф с порогом $d = 1.0$ и вычислялись четыре признака.

4.2 Анализ важности признаков

При помощи RandomForest оценивалась важность признаков при $n = 25, 100, 500$. Результаты приведены в таблице:

Признак	$n = 25$	$n = 100$	$n = 500$
count_triangles	0.49	0.45	0.45
clique_number	0.34	0.39	0.39
min_degree	0.00	0.01	0.05
connected_components	0.16	0.15	0.11

Таблица 5: Важность признаков при разных размерах выборки

Вывод: `count_triangles` и `clique_number` являются наиболее информативными.

4.3 Классификация и метрики качества

Эксперименты проводились для $n = 10, 20, 50, 100, 200, 500$ с классификаторами LogisticRegression, RandomForest и SVM. Оценивались Accuracy, дисперсия Accuracy, FPR, TPR, Precision и F1.

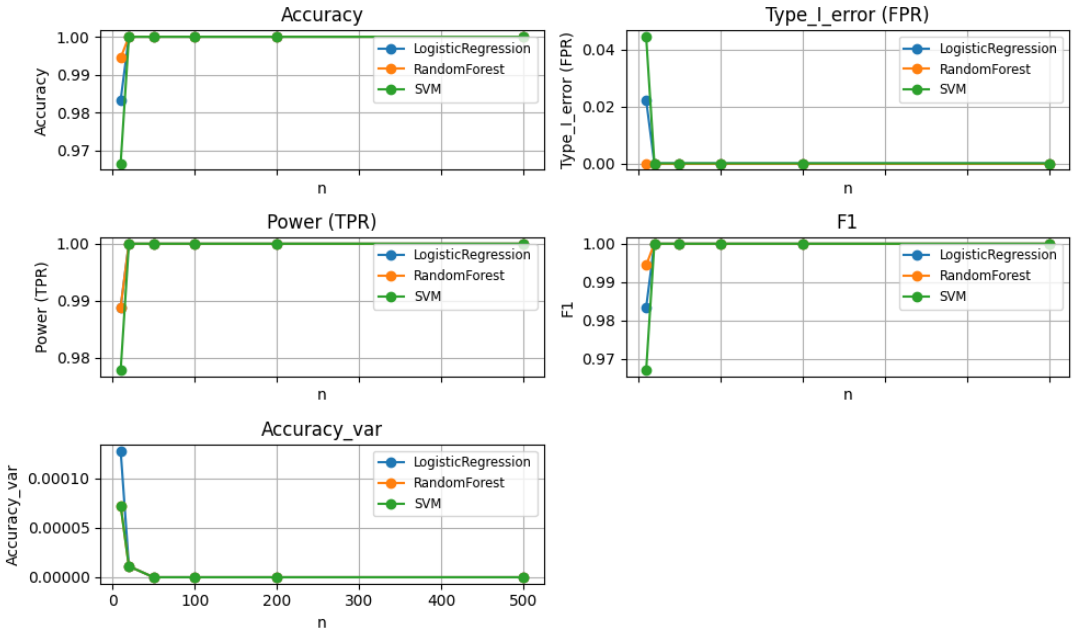

Рис. 3: Зависимость метрик качества от размера выборки

5 Выводы (Часть II)

- При $n \geq 20$ все алгоритмы достигают 100% Accuracy и мощности, при этом FPR $= 0$.
- Для практических задач достаточно $n \approx$ 20-50 для идеального разделения.
- RandomForest и SVM показали наилучшую стабильность при малых выборках.
- Наиболее информативные признаки: `count_triangles` и `clique_number`.