Отчет по части II

Богданов Ярослав

Гипотезы:

- H_0: данные из распределения weibull с параметром $\lambda = 1$
- H_1: данные из распределения lognormal с параметром $\sigma = log(5)$

Параметры исследования:

- Тип графа: dist-граф с параметром $d = 0.5$
- Размеры выборок: $n = 25, 100, 500$
- Количество выборок на класс: 100

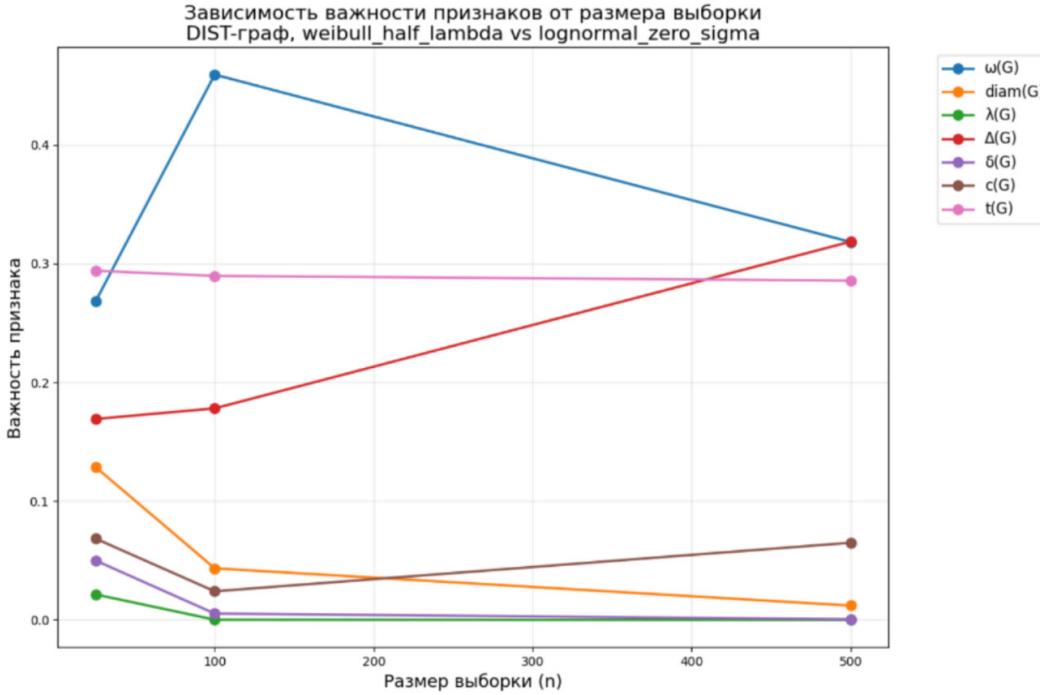
Исследуемые характеристики графов:

- $\Delta(G)$ — максимальная степень вершины
- $\delta(G)$ — минимальная степень вершины
- $c(G)$ — количество компонент связности
- $t(G)$ — количество треугольников
- $\mathrm{diam}(G)$ — диаметр графа
- $\lambda(G)$ — рёберная связность
- $\omega(G)$ — кликовое число

1 Результаты

1.1 Анализ важности характеристик

Анализ важности характеристик с использованием Random Forest показал следующие результаты:



Основные наблюдения:

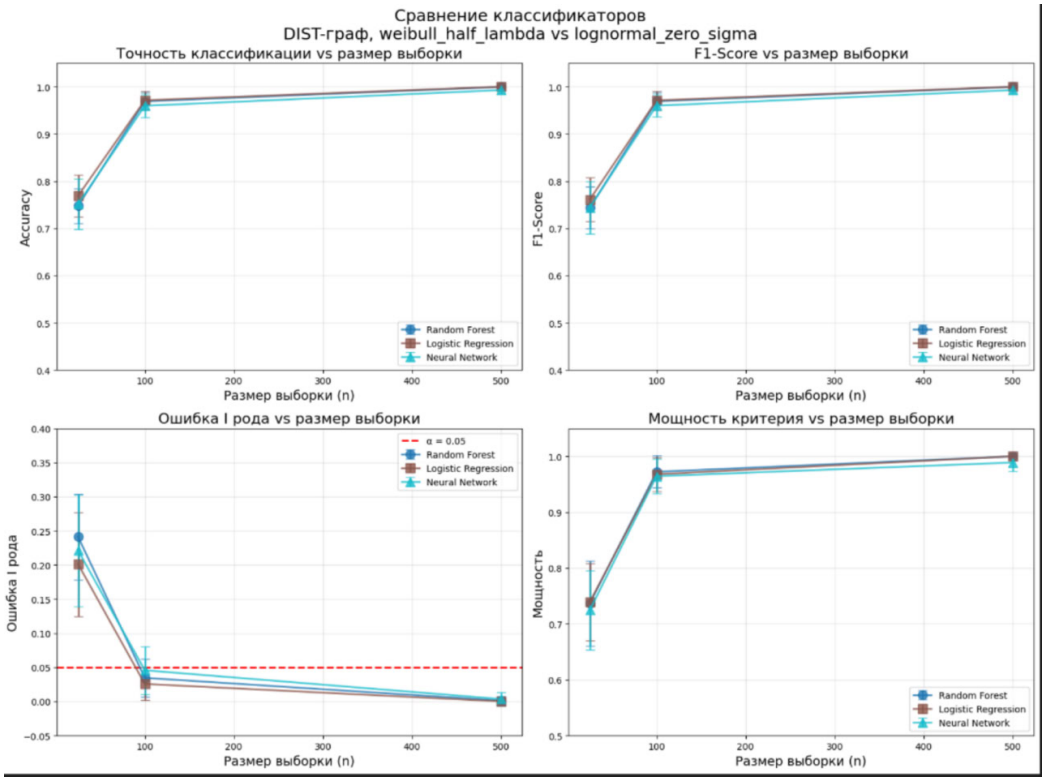
- При $n = 25$ наибольший вклад дают число треугольников $t(G) \approx 29\%$ и кликовое число $\omega(G) \approx 27\%$, за ними следуют максимальная степень $\Delta(G) \approx 17\%$ и диаметр $\mathrm{diam}(G) \approx 13\%$.
- При $n = 100$ доминирует $\omega(G) \approx 46\%$, тогда как важность диаметра $\mathrm{diam}(G)$ падает до $\approx 4\%$, а $\delta(G)$ и $\lambda(G)$ практически сходят на нет ($< 1\%$).
- При увеличении до $n = 500$ максимальная степень $\Delta(G)$ растёт до $\approx 32\%$ и выравнивается с $\omega(G) \approx 32\%$, число треугольников $t(G)$ остаётся стабильным ($\approx 29\%$).
- Диаметр $\mathrm{diam}(G)$ продолжает снижаться (до $\approx 1.3\%$), а минимальная степень $\delta(G)$ и реберная связность $\lambda(G)$ практически теряют

значение.

- Число компонент связности $c(G)$ демонстрирует U-образную динамику: $\approx 7\% \to 2\% \to 6.5\%$ при росте n.

1.2 Сравнение классификаторов

Для оценки качества классификации использовались следующие алгоритмы: Random Forest, Logistic Regression и Neural Network. Результаты представлены на графике:

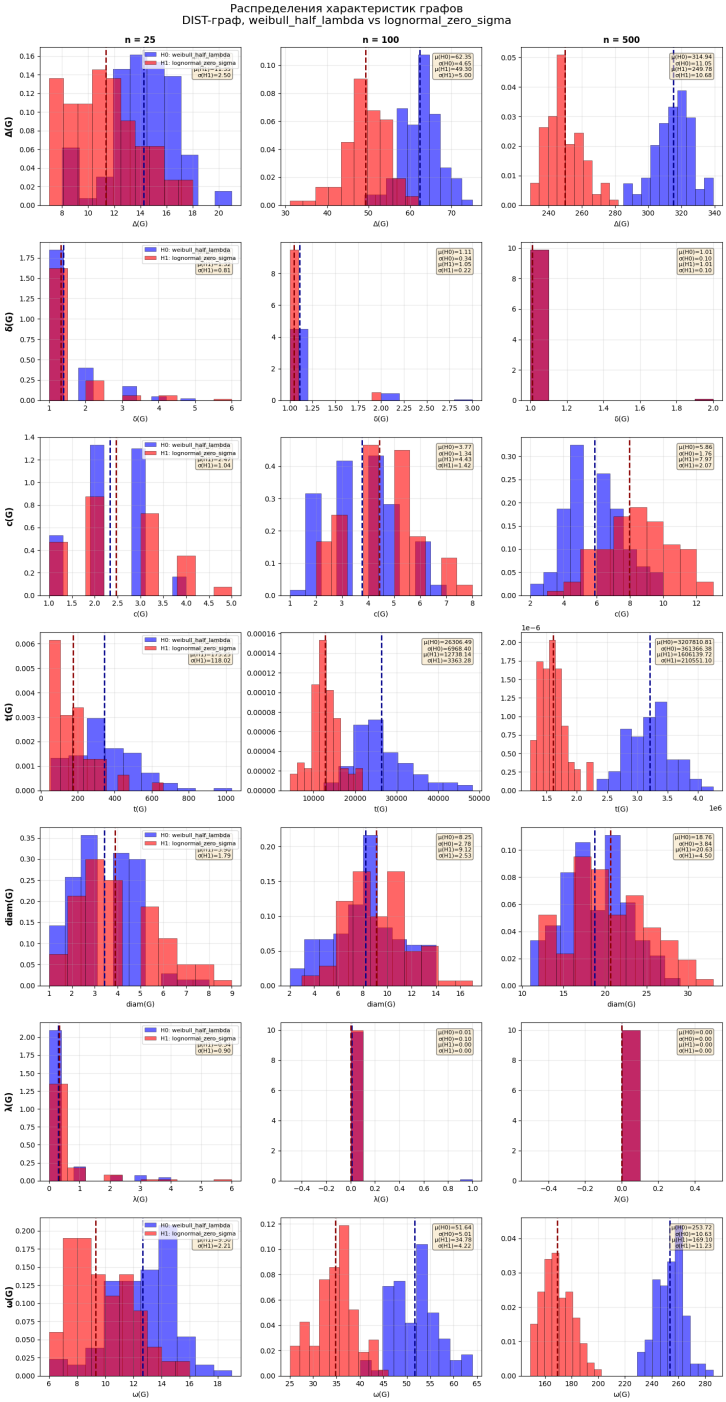


Основные выводы по классификаторам:

- Для малых выборок ($n = 25$) все модели демонстрируют среднюю точность (≈ 0.75–0.77) и аналогичный F1-Score, при этом ошибка I рода значительно превышает уровень значимости ($\alpha = 0.05$), достигая 20%–24%, а мощность критерия находится на уровне 0.72–0.74.
- При увеличении выборки до $n = 100$ точность и F1-Score резко возрастают до 0.95–0.97, ошибка I рода падает ниже 5% (до $\approx$ 2%–3%), а мощность критерия достигает 0.94–0.97.

- Для больших выборок ($n = 500$) все три алгоритма достигают практически идеальных показателей: точность и F1-Score близки к 1.00, ошибка I рода стремится к нулю, мощность критерия приближается к единице.
- Различия между алгоритмами минимальны: Logistic Regression чуть опережает Random Forest на средних выборках, Neural Network демонстрирует чуть больший разброс оценок.

1.3 Анализ распределений характеристик



Гистограммы распределений характеристик графов показывают, как изменяется разделимость между гипотезами H_0 и H_1 с ростом размера выборки:

- $\Delta(G)$ (максимальная степень) — при $n = 25$ видна лишь слабая тенденция к сдвигу, при $n = 100$ распределения уже хорошо разделяются, а при $n = 500$ их разделение становится почти полным.
- $t(G)$ (число треугольников) — умеренное разделение для $n = 25$ и $n = 100$, для $n = 500$ гистограммы практически не перекрываются.
- $\mathrm{diam}(G)$ (диаметр) — заметное, но неполное разделение; с ростом n средние значения расходятся, но хвосты всё ещё пересекаются.
- $\omega(G)$ (кликовое число) — при $n = 100$ уже явное разделение, при $n = 500$ гистограммы хорошо разделены.
- $c(G)$ (число компонент связности) — небольшое смещение средних при $n \geq 100$, сильнее выраженное при $n = 500$, но перекрытие сохраняется.
- $\delta(G)$ (минимальная степень) и $\lambda(G)$ (рёберная связность) — при любых n распределения почти совпадают, разделения не наблюдается.

С увеличением размера выборки разделение между распределениями становится более выраженным, что объясняет улучшение качества классификации на больших n.

2 Выводы

Анализ итоговых показателей классификации даёт следующие выводы:

- **Для** $n = 25$: ни один классификатор не удовлетворяет условию $\alpha \leq 0.05$.
- **Для** $n = 100$: лучший классификатор — Random Forest с ошибкой I рода $\alpha = 0.0344$ и мощностью 0.9722.
- **Для** $n = 500$: лучший классификатор — Random Forest с ошибкой I рода $\alpha = 0.0011$ и мощностью 1.0000.