Отчет по части II

Шаяхметов Аскар

Гипотезы:

- H_0: данные из распределения skewnorm с параметром $\alpha = 1$
- H_1: данные из распределения student_t с параметром $\nu = 3$

Параметры исследования:

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

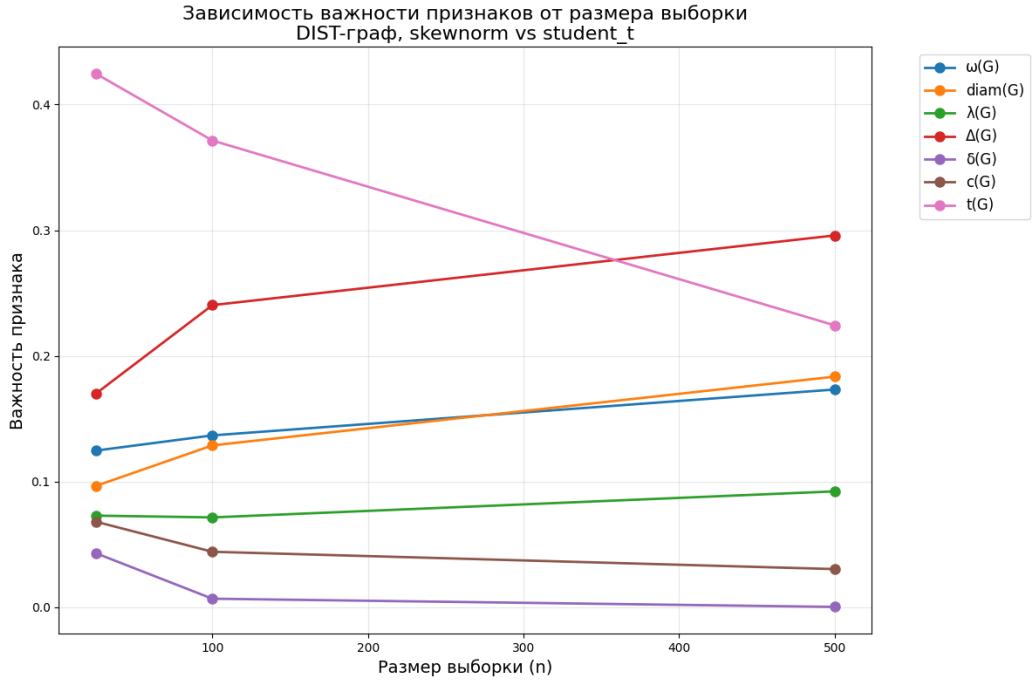
Исследуемые характеристики графов:

- $\Delta(G)$ — 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:

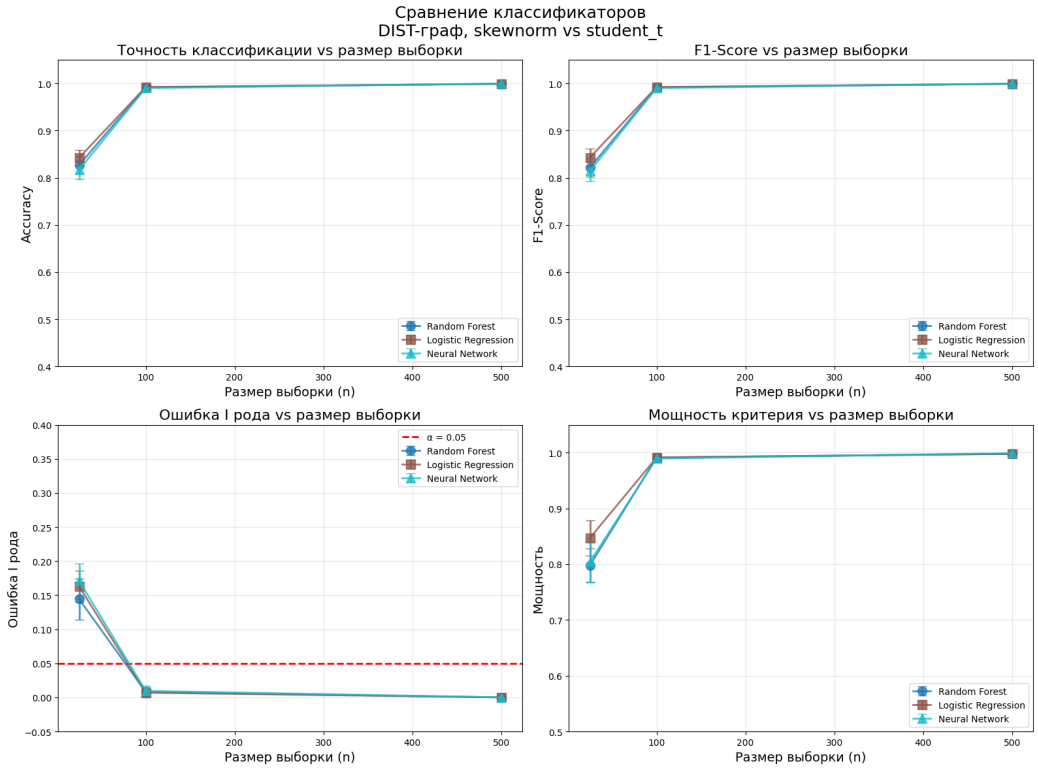


Основные наблюдения:

- Для малых выборок ($n = 25$) наиболее важной характеристикой является количество треугольников $t(G)$ (42.5% важности)
- С ростом размера выборки важность максимальной степени $\Delta(G)$ увеличивается: от 17% при $n = 25$ до 29.6% при $n = 500$
- Минимальная степень $\delta(G)$ практически теряет значение с ростом n

1.2 Сравнение классификаторов

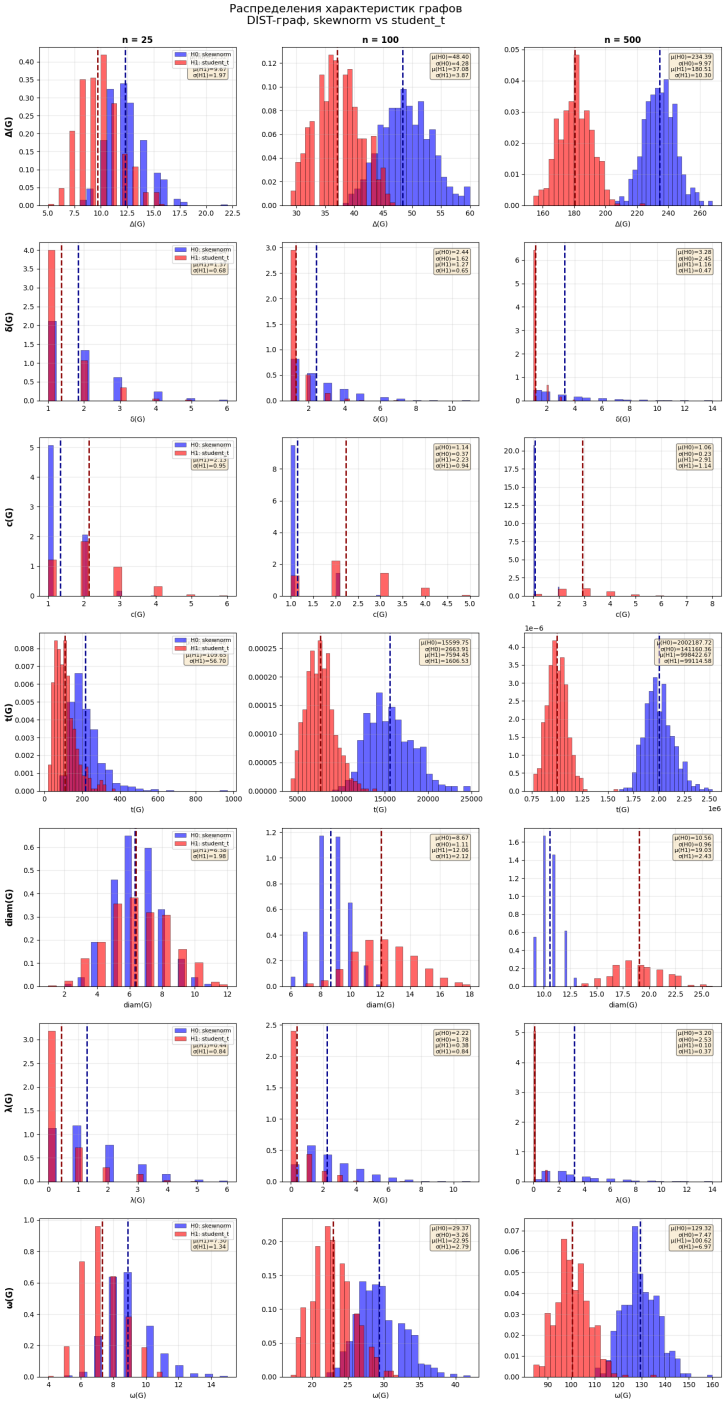
Для оценки качества классификации использовались следующие алгоритмы: Random Forest, Logistic Regression и Neural Network. Результаты представлены на графике:



Основные выводы по классификации:

- Для малых выборок ($n = 25$) все классификаторы показывают умеренное (≈ 0.83) качество с высокой ошибкой первого рода ($\alpha > 0.14$)
- При $n = 100$ качество классификации резко улучшается, ошибка первого рода снижается до уровня ($\alpha \approx 0.01$)
- Для больших выборок ($n = 500$) все классификаторы показывают практически идеальное качество

1.3 Анализ распределений характеристик



Гистограммы распределений характеристик графов показывают четкое разделение между гипотезами H_0 и H_1 для некторых характеристик.

- Максимальной степени $\Delta(G)$ — разделение улучшается при увеличении n
- Количества треугольников $t(G)$ — четкое разделение для $n = 500$
- Диаметра графа $\mathrm{diam}(G)$ — приемлемое разделение
- Кликового числа $\omega(G)$ — для $n = 500$ хорошее разделение

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

2 Выводы

Анализ результатов показал следующее:

- **Для** $n = 25$: ни один классификатор не удовлетворяет условию $\alpha \leq 0.05$
- **Для** $n = 100$: лучший классификатор — Random Forest с ошибкой первого рода $\alpha = 0.008$ и мощностью 0.991
- **Для** $n = 500$: лучший классификатор — Neural Network (два скрытых слоя размерами 50 и 30) с ошибкой первого рода $\alpha = 0.000$ и мощностью 0.999