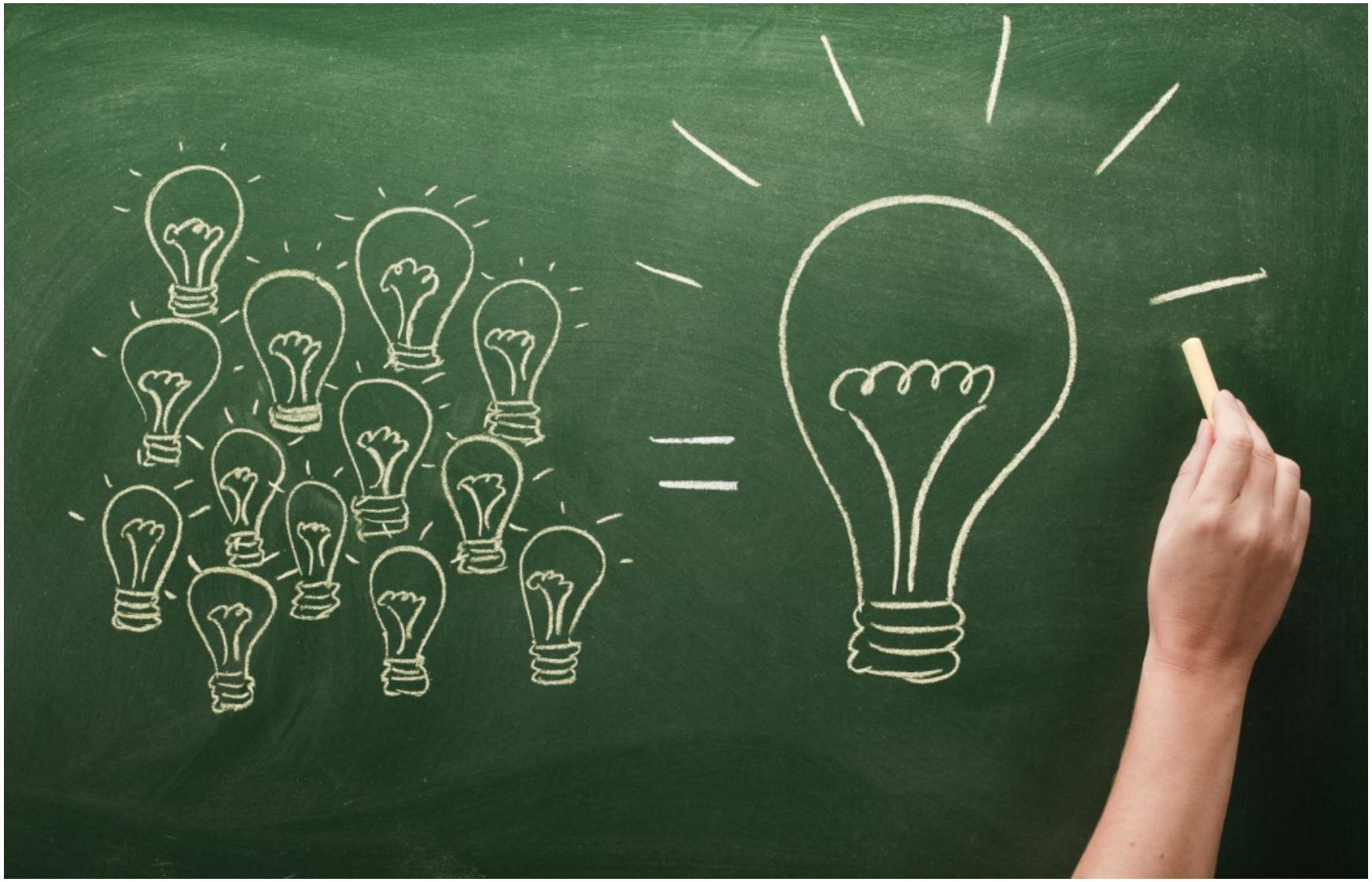
АЛГОРИТМ ROTATION FOREST : СРАВНИЕТЛЬНЫЙ АНАЛИЗ С ИЗВЕСТНЫМИ АНСАМБЛЯМИ КЛАССИФИКАТОРОВ

PANIN KIRILL

FACULTY OF APPLIED MATHEMATICS AND COMPUTER SCIENCE OF THE BELARUSIAN STATE UNIVERSITY

Power of the crowds. Wisdom of the crowds



Задачи исследования

- **Оценить тенденцию изменений эффективности классификации с использованием критерия точности классификации для ансамблей с увеличением из размерности (количества деревьев). Разработать способ обобщения оценок точности классификации ансамблями, полученных по всем наборам данных.**
- **Выполнить сравнительный анализ эффективности классификации для ансамблей с фиксированным размером по всем наборам данных. Особый интерес представляют ансамбли малой размерности, т.к. известно, что эффективности различных методов ансамблевой классификации сравниваются с увеличением числа деревьев. Для сравнения разработать способ ранжирования ансамблей на основе попарного сравнения эффективности по совокупности всех анализируемых наборов данных.**
- **Оценить отличие эффективности предложенного метода на основе вращающихся деревьев от других ансамблей классификаторов с использованием как табличного, так и графического представления**

Эксперименты

Постановка эксперимента

Ансамбли классификаторов:

- Метод бэггинг
- Метод бустинг
- Метод случайных лесов
- Предложенный метод на основе вращающихся лесов

Базовый алгоритм:

- Деревья решений CART

Наборы данных:

- 14 наборов данных из UCI Machine Learning Repository и KEEL-dataset repository

Table 1
Характеристики наборов данных использованных в исследовании

Имя датасета	Атрибуты (R/I/N)	Количество объектов	Количество классов
Appendicitis	7 (7/0/0)	106	2
Balance	4 (4/0/0)	625	3
BreastCancer	5 (5/0/0)	215	3
Bupa	6 (1/5/0)	345	2
Cleveland	13 (13/0/0)	297 (303)	5
Ecoli	7 (7/0/0)	336	8
Heart	13 (1/12/0)	270	2
Ionosphere	33 (32/1/0)	351	2
Iris	4 (4/0/0)	150	3
Led7digit	7 (7/0/0)	500	10
Pima	8 (8/0/0)	768	2
Sonar	60 (60/0/0)	208	2
Vehicle	18 (0/18/0)	846	4
Wine	13 (13/0/0)	178	3

Experimental results

Постановка эксперимента

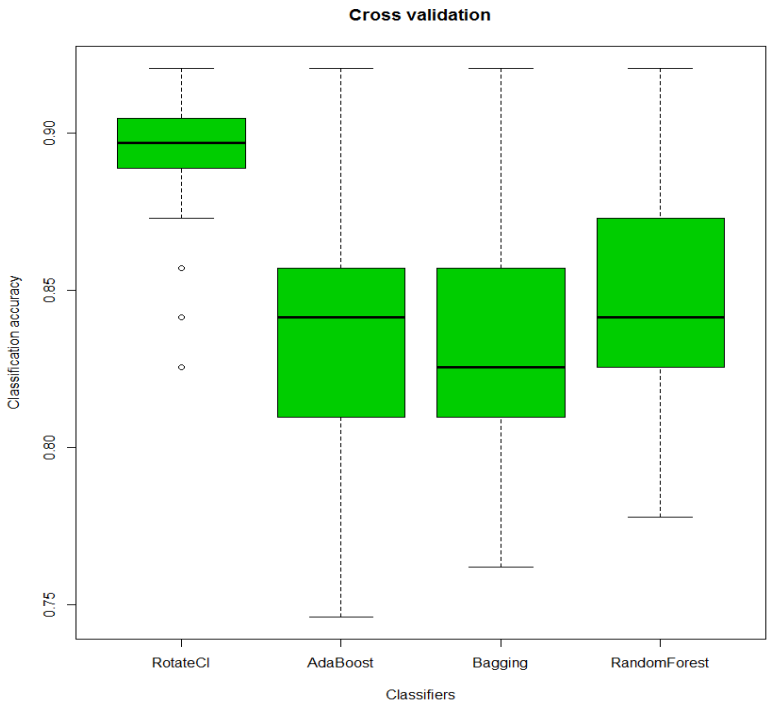
- Размер ансамблей варьировался от 10 до 100 с шагом 10. Эффективность классификации сравнивалась для ансамблей равных размеров. Интерес представляют ансамбли малых фиксированных размеров ансамблей.
- Для каждого набора данных и ансамбля эффективность классификации оценивалась с использованием 100 повторных разбиений объектов на обучающее и тестовое множества в пропорции 10:1. Т.е. ансамбль строился на обучающем множестве, а эффективность классификации оценивалась на тестовом. Данный подход позволил оценить среднее значение и стандартное отклонение критерия эффективности классификации.
- В качестве критерия оценки эффективности классификации рассчитывалась точность классификации и результаты для каждого ансамбля представляются.

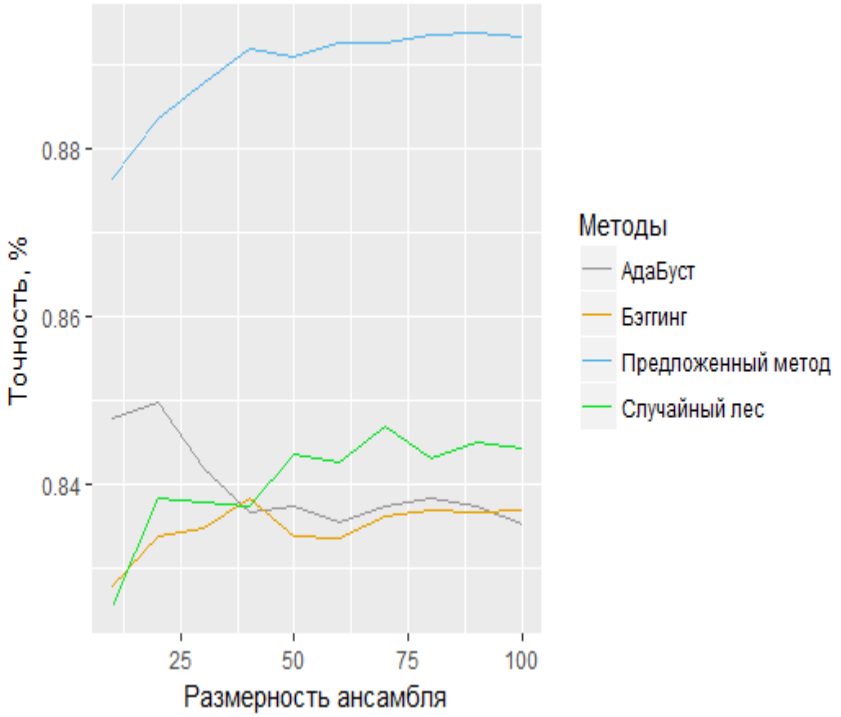
Table 2

Точность классификации(Classification Accuracy) и стандантное отклонение(Standard Deviation)

datasets	Rotation Forest	AdaBoost	Bagging	Random Forest
Appendicitis	0.848667±0.003338	0.84275±0.003168	0.84725±0.00358	0.85725±0.005458 •
•Balance	0.889524±0.005594 •	0.839778±0.005067	0.834968±0.002947	0.840444±0.006132
BreastCancer	0.966521±0.000835 •	0.960676±0.001713	0.957014±0.00154	0.96607±0.002032
Bupa	0.706114±0.005623	0.709971±0.006335	0.720229±0.008857	0.726371±0.015135 •
Cleveland	0.554594±0.003517 •	0.52175±0.006212	0.535187±0.002201	0.539062±0.004459
Ecoli	0.803395±0.005859	0.784947±0.003797	0.762763±0.002867	0.805684±0.007071 •
•Heart	0.846963±0.003768 •	0.809556±0.008433	0.818815±0.00618	0.826222±0.010765
•Ionosphere	0.954167±0.00369 •	0.93775±0.001759	0.921222±0.002706	0.933167±0.002597
Iris	0.947733±0.002935	0.946±0.002534	0.953533±0.002704 •	0.950467±0.001635
Led7digit	0.726182±0.003557	0.730964±0.001934 •	0.719582±0.002574	0.721636±0.00698
Pima	0.770078±0.002707 •	0.743±0.003503	0.76974±0.002654	0.763052±0.006185
Sonar	0.835591±0.014648 •	0.829±0.020402	0.780091±0.004762	0.817727±0.016513

Набор данных:
Balance.





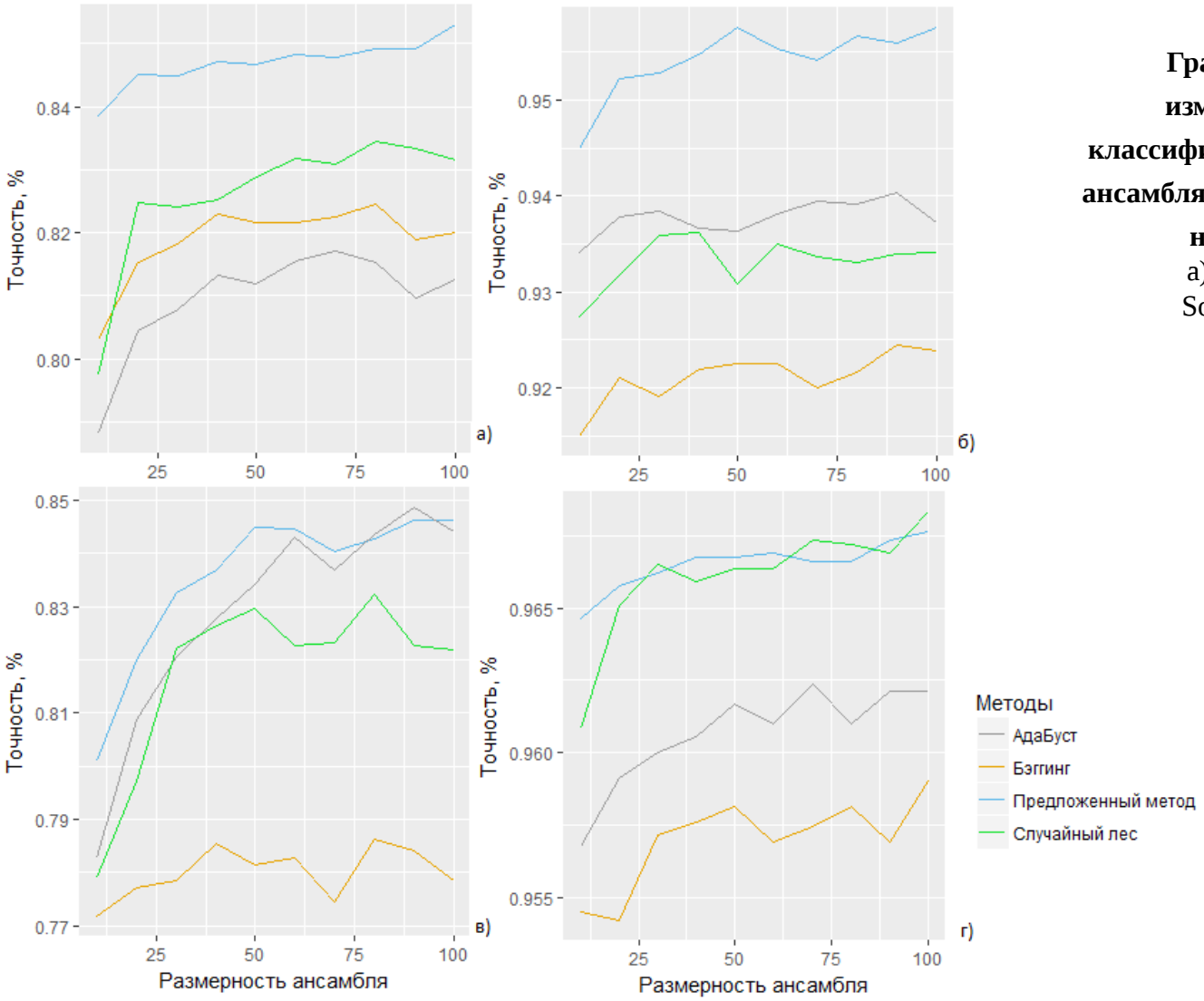


График демонстрирует изменение точности классификации от размерности ансамбля для каждого метода на наборе данных:
а) Heart, б) Ionosphere, в) Sonar, г) Breast Cancer

Ранжированный список сравниваемых методов, согласно разнице между числом раз, когда метод был лучше или хуже другого метода при проведении попарных сравнений.

	RotationForest	AdaBoost	Bagging	RandomForest
Rank	20	-12	-22	14
Wins	31	15	10	28
Losses	11	27	32	14

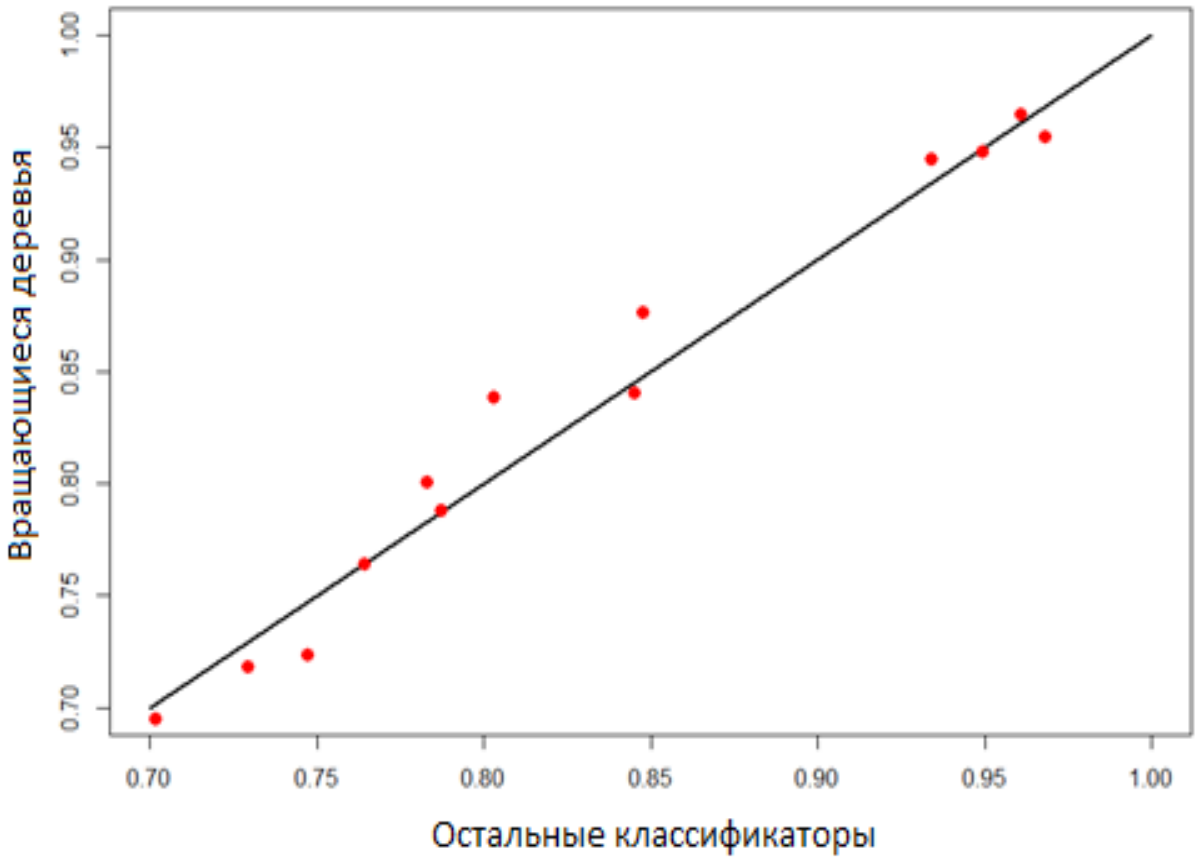


Диаграмма демонстрирует точность предложенного метода к остальным методам по всем наборам данных

На оси Y представлена точность предложенного метода, а на оси X наилучшая точность классификации среди сравниваемых методов. Диагональной линия, соответствует одинаковым значениям точности. Большинство точек лежит выше диагональной линии означает преимущество предложенного метода.

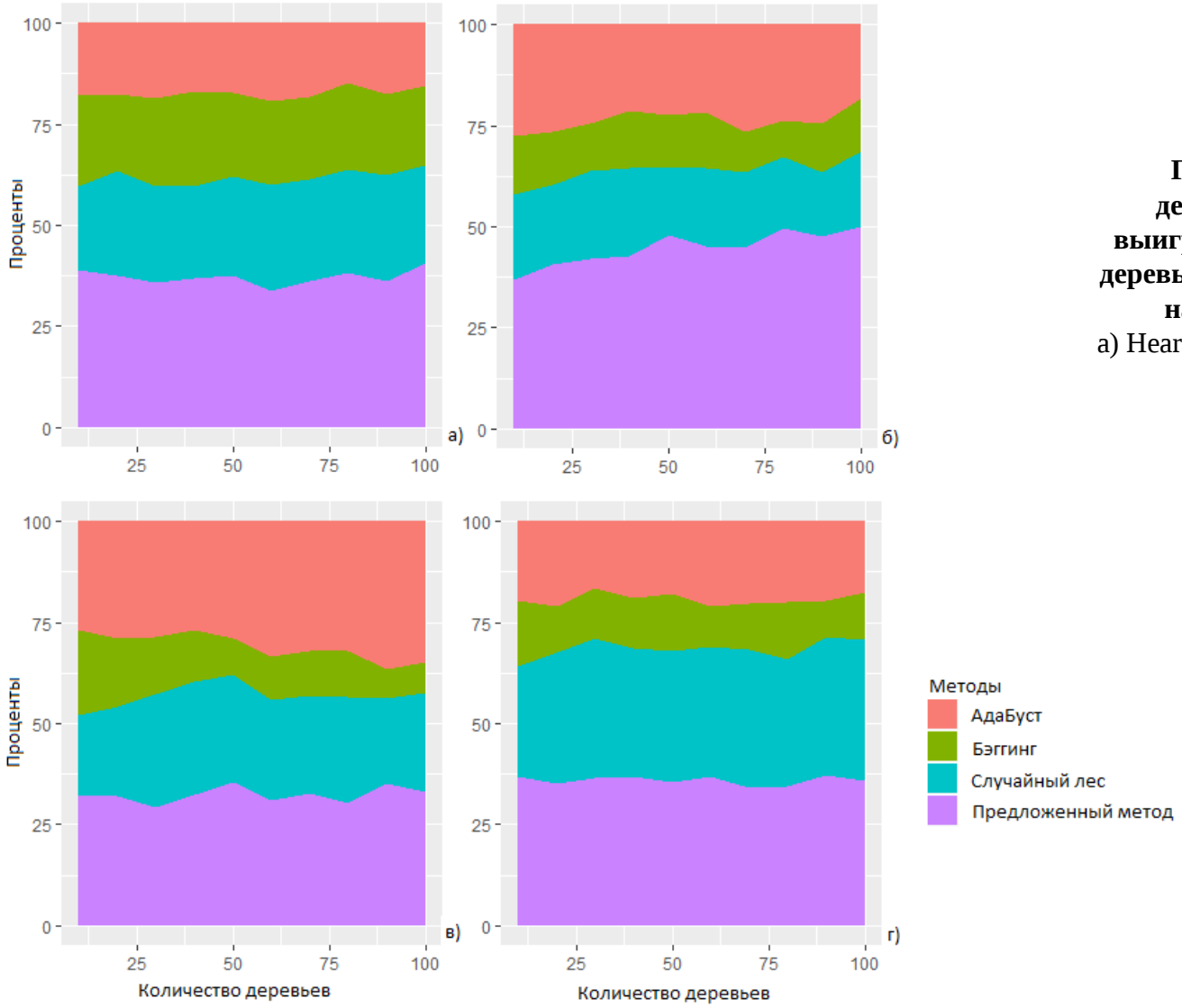
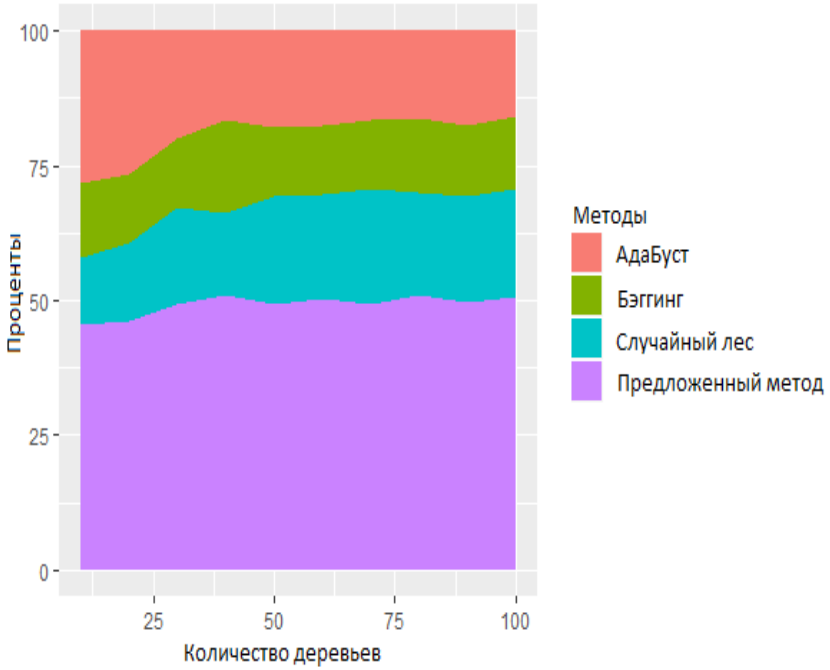


График со слоями демонстрирует долю выигрышей от количества деревьев для каждого метода на наборе данных :
а) Heart, б) Ionosphere, в) Sonar, г) Breast Cancer



.График со слоями демонстрирует долю выигрышей для каждого метода на наборе данных «Balance»

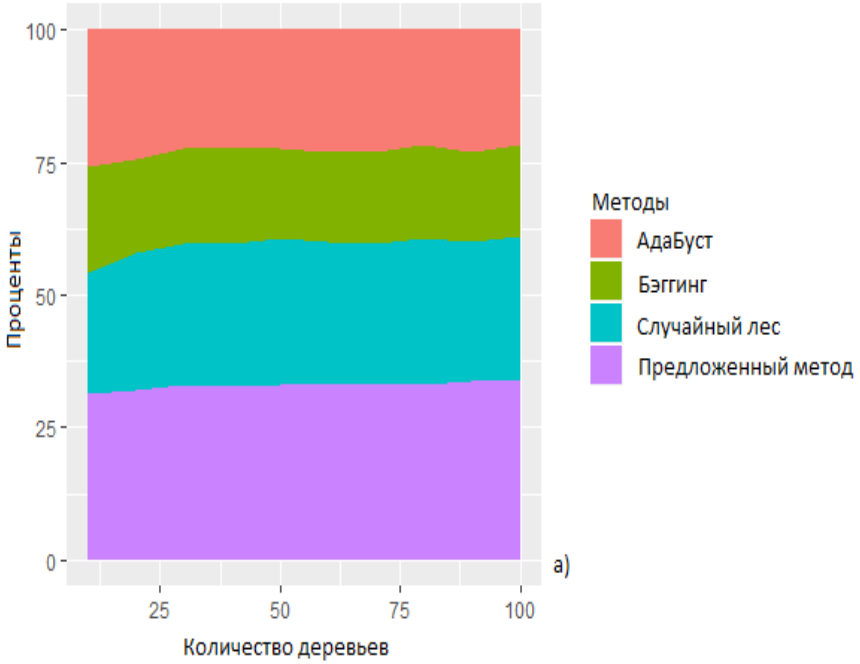


График со слоями демонстрирует долю выигрышей для каждого метода на всех наборах данных в зависимости от количества деревьев

Выводы

В ходе исследований проведен сравнительный анализ нескольких ансамблей классификаторов, включая предложенную ансамбль на основе вращающихся деревьев. Гипотеза о преимуществах предложенного ансамбля подтвердилась на большинстве проанализированных наборах данных. Кроме того получены следующие результаты:

- Проанализированы изменения эффективности классификации с использованием критерия точности классификации для ансамблей с увеличением их размерности. Разработан способ обобщения оценок точности классификации ансамблями, полученных по всем наборам данных. Для визуализации результатов классификации разработан вариант графического представления эффективности ансамблей на наборах данных в виде процентной диаграммы.
- Выполнен сравнительный анализ эффективности классификации для ансамблей с фиксированным размером по всем наборам данных. Разработан способ ранжирования ансамблей на основе попарного сравнения эффективности по совокупности всех анализируемых наборов данных.
- С помощью разработанного способа ранжирования ансамблей на основе попарного сравнения эффективности по совокупности всех анализируемых наборов данных получили, что метод вращающихся деревьев стоит на первом месте в списке ансамблей классификаторов, ранжированном по эффективности. На втором месте стоит случайный лес. При анализе эффективности ансамблей по совокупности наборов данных

Thank you!