Методы ансамблирования обучающихся алгоритмов

Гущин Александр

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель: д.ф.-м.н. А. Г. Дьяконов

Группа 174, весна 2015

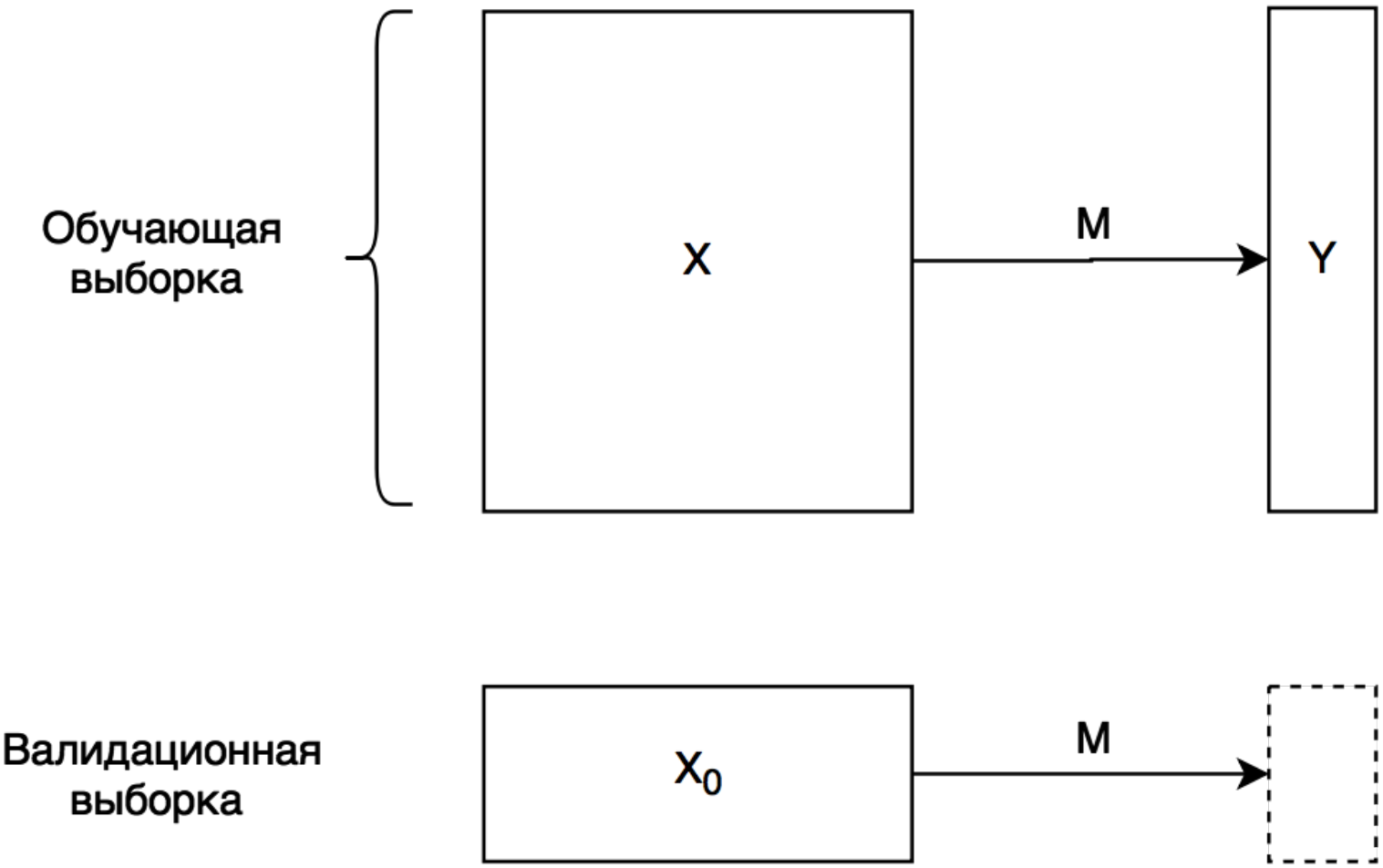
Цель исследования

Цель работы: Стекинг (Stacking) - один из наиболее эффективных в смысле достигаемого качества способов ансамблирования обучающихся алгоритмов. Цель работы - предложить новую модификацию алгоритма стекинга и экспериментально сравнить её с предложенными ранее.

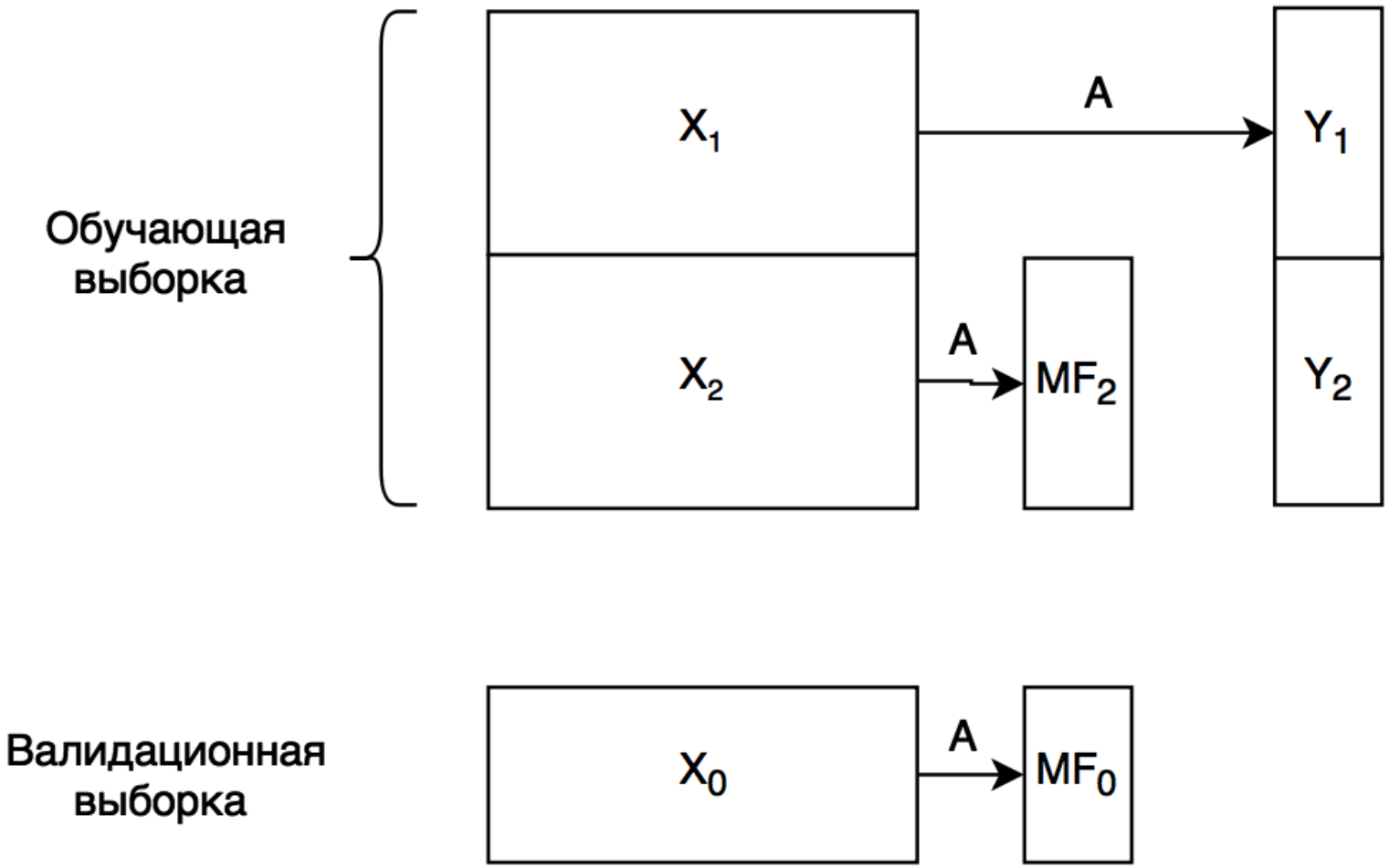
Решение:

- разработать алгоритм
- поставить вычислительные эксперименты

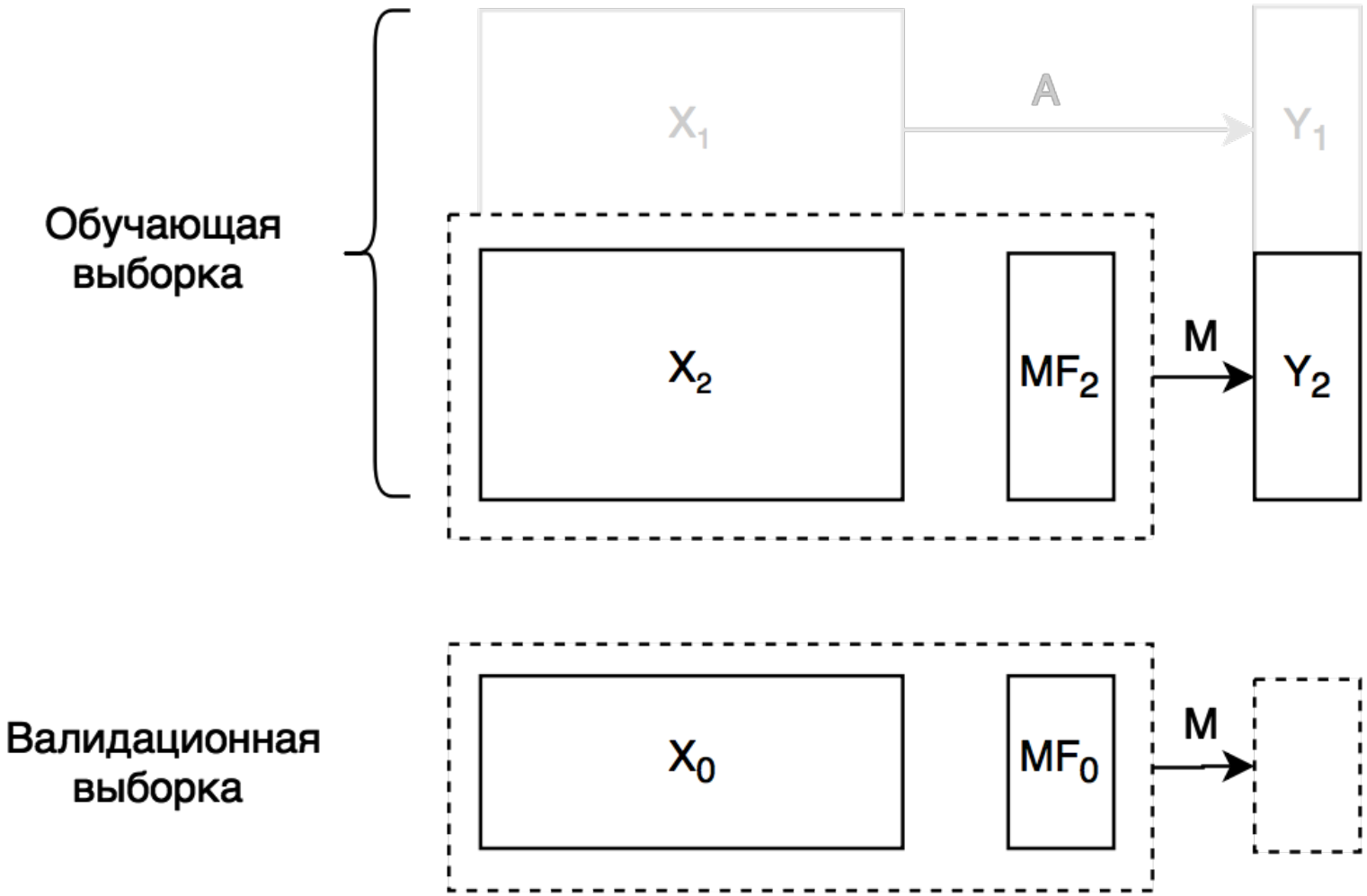
Постановка задачи



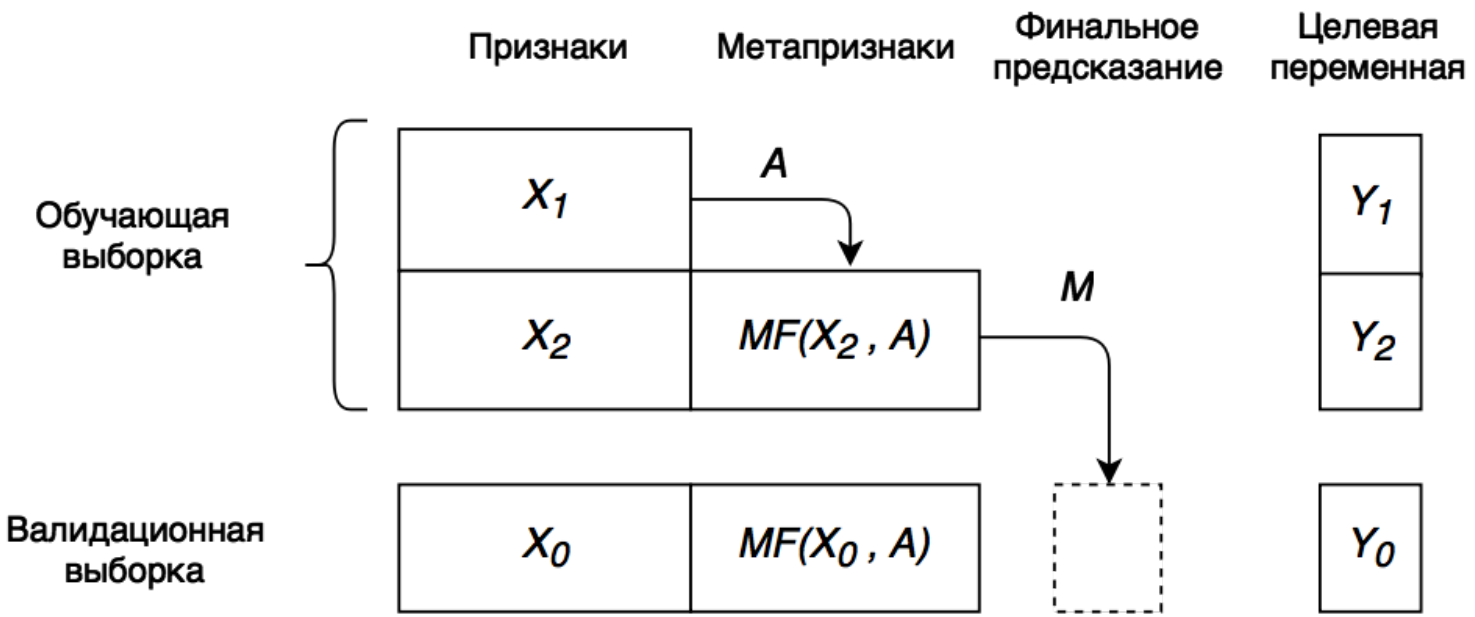
Стекинг: шаг 1



Стекинг: шаг 2

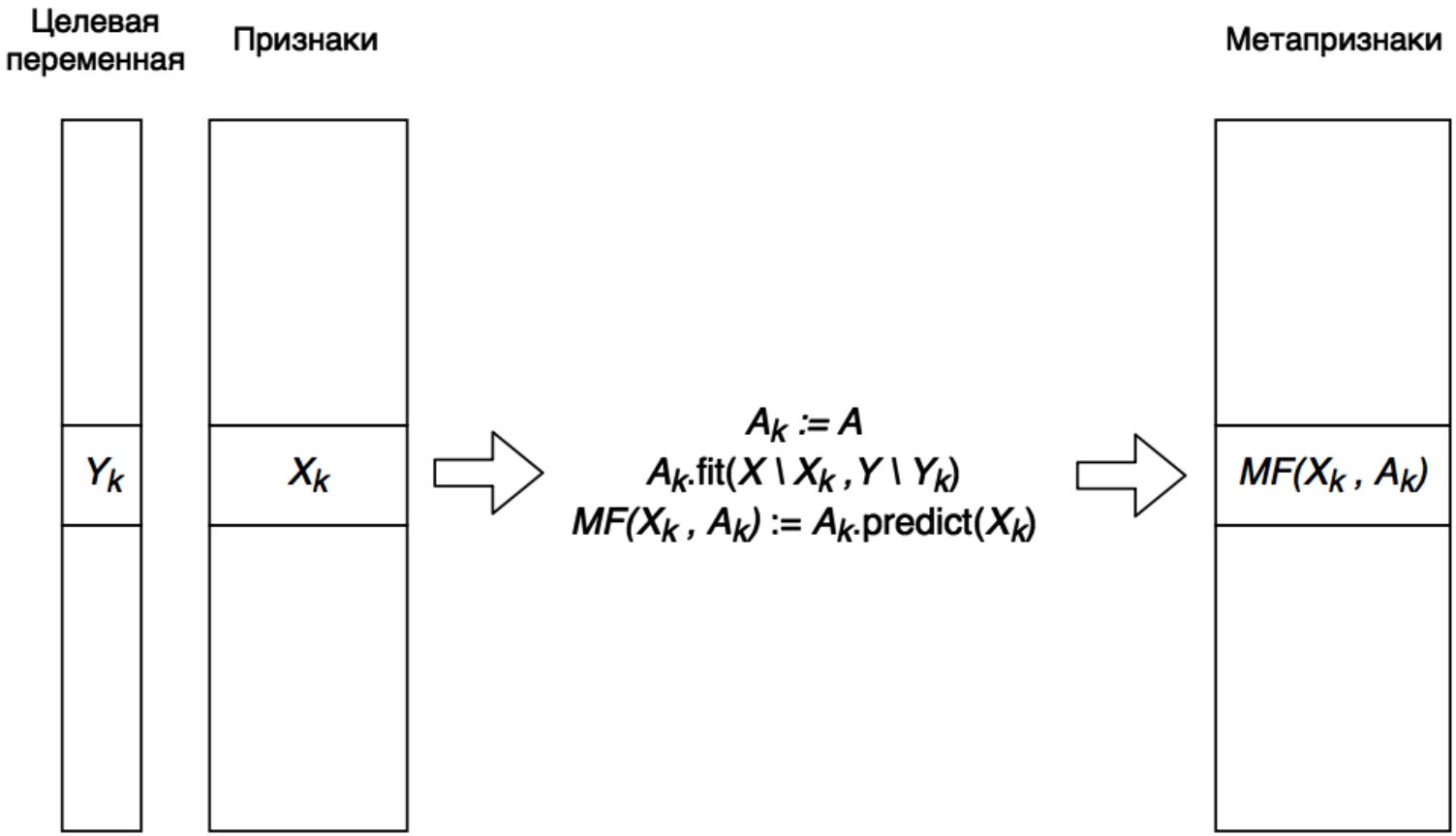


Стекинг по схеме hold-out

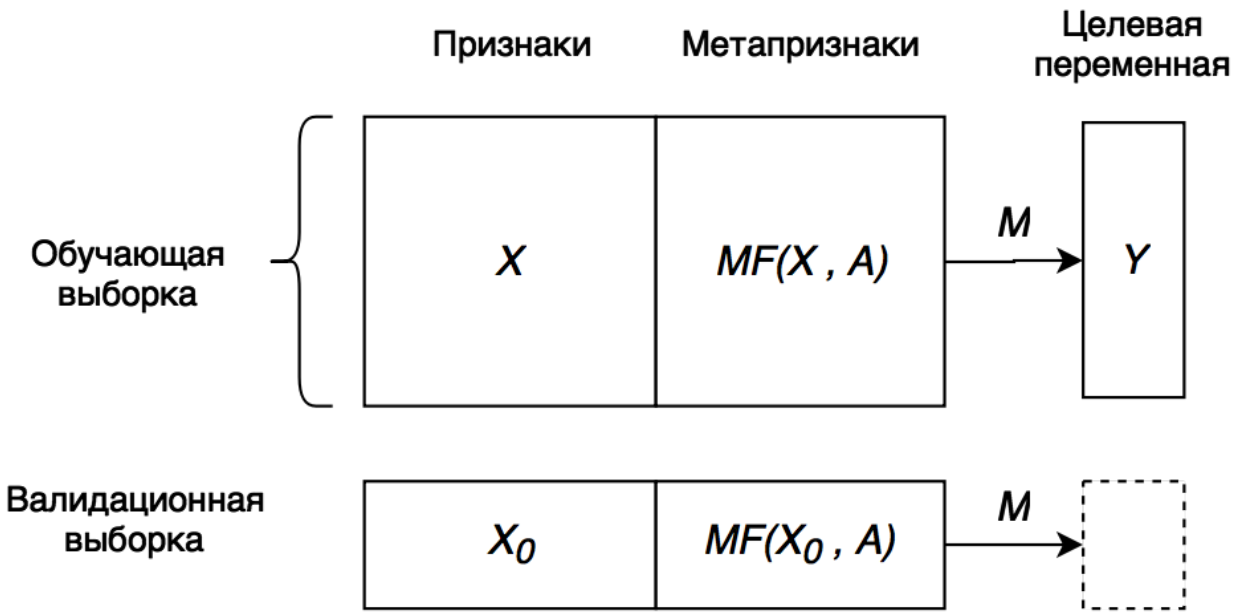


Преимущества: отсутствие переобучения
Недостатки: неэффективное использование обучающей выборки

Получение метапризнака по схеме K-fold

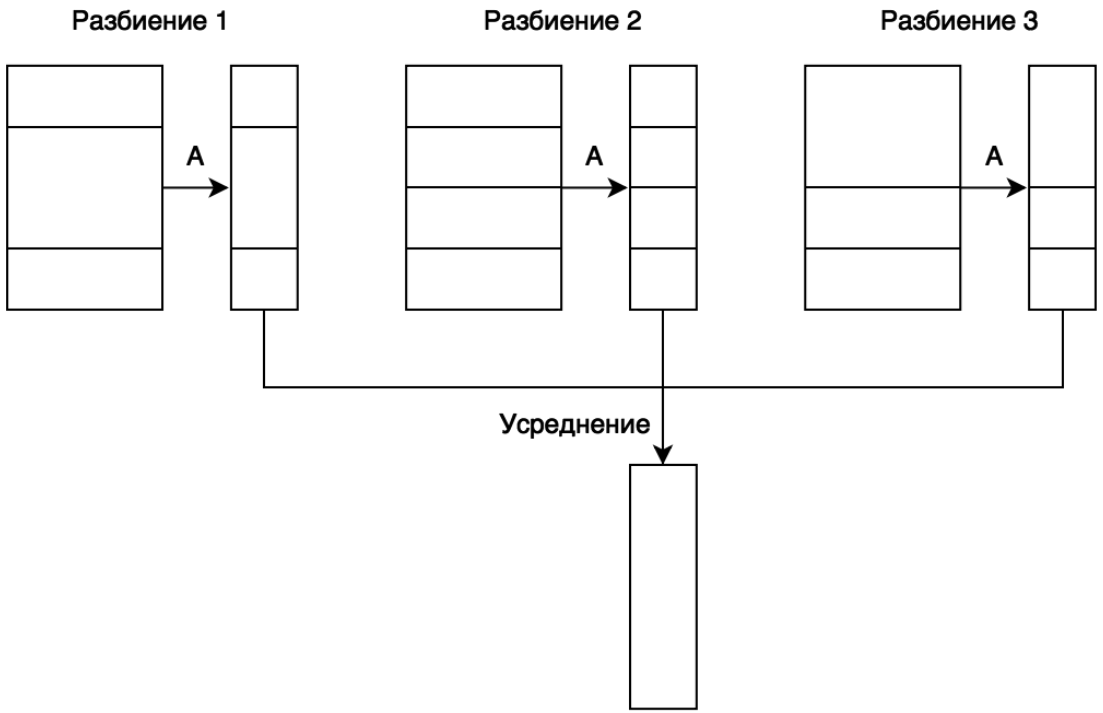


Стекинг по схеме K-fold



Преимущества: использование всей обучающей выборки
Недостатки: переобучение, "неоднородность"метапризнаков

Идея предлагаемой модификации



Преимущества: эффективное использование обучающей выборки, борьба с "неоднородностью"метапризнаков

Недостатки: увеличенная вычислительная сложность

Поставленные эксперименты

Данные:

- UCI, Forest Cover Type Prediction;
- Kaggle, Otto Group Product Classification Challenge.

$|\text{Train}| \approx |\text{Test}| \approx 15000$.

Все результаты приводятся для усреднения по 10 случайным разбиениям на обучающую и тестовую выборку.

Выдвигаемые предположения

- метапризнаки, полученные с помощью усреднений имеют качество лучше, чем метапризнаки, полученные без усреднений
- улучшение непосредственного качества метапризнака, связанное с выбором способа его построения, не обязательно ведёт к улучшению качества метаклассификатора

Качество метапризнаков, полученных различными способами

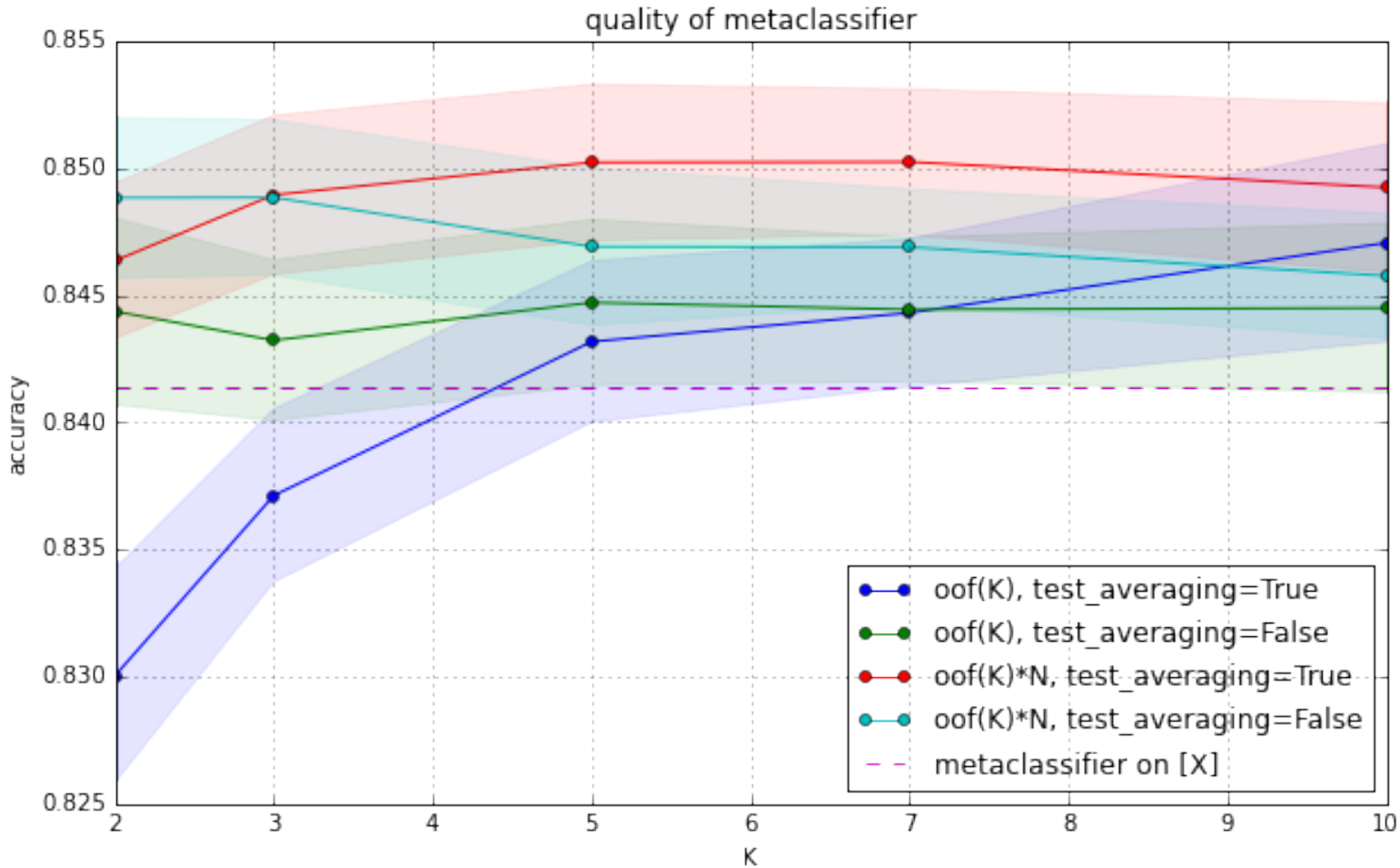
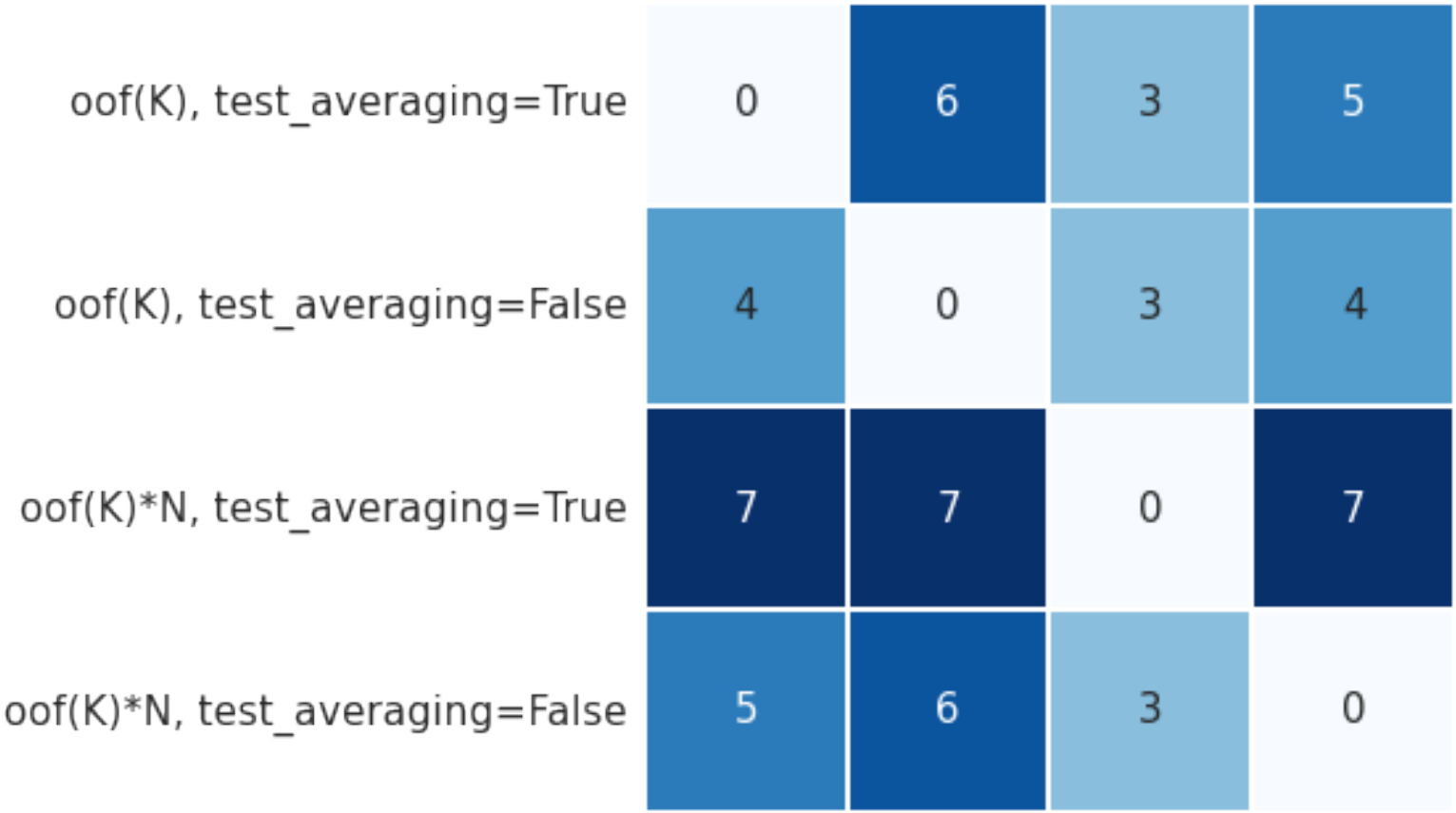


Рис.: ExtraTreesClassifier(X, XGBoost)

Результаты экспериментов



Результаты, выносимые на защиту

- Предложена модификация алгоритма построения метапризнаков
- В поставленых экспериментах продемонстрировано, что предложенная модификация часто оказывается лучше предложенной ранее