Использование алгоритмов машинного обучения для определения неэффективно кэшируемой нагрузки

Киргизов Григорий, 344 гр.
Научный руководитель:
руководитель исследовательской лаборатории Raidix, к.т.н. Лазарева С. В.

Введение: гибридные СХД

- Гибридные СХД используют SSD в качества кэша
- Традиционные алгоритмы кэширования (LRU, LFU) не предназначены для SSD
  - SSD быстро изнашиваются

Может, не кэшировать “ненужные” данные?

Подход: машинное обучение

- Есть много данных — истории запросов к СХД
- Попробуем по историям прогнозировать эффективность данных для кэширования

Глубокое обучение:

- Успешно применяется в анализе речи и текста
- Наши данные имеют схожую природу
  - Временные ряды со сложными зависимостями
- Попробуем применить

Постановка задачи

Цель работы:

- Исследовать применимость методов машинного обучения к проблеме кэширования в гибридных СХД

Задачи:

- Провести обзор подходящих методов машинного обучения
- Составить план экспериментов с учетом обзора
- Подготовить данные для их проведения
- Провести эксперименты
- Выполнить анализ результатов

Обзор методов машинного обучения

- Машинное обучение:
  - Ансамбли деревьев решений (Gradient Tree Boosting - XGBoost [2016] library)
- Глубокое обучение:
  - LSTM Recurrent Neural Networks (RNN) [2000]
  - Pre-training + LSTM RNN [2015]
  - Hierarchical RNN [2015]
  - RNN with Skip Connections [2016]

Описание работы

Данные

Метрика:

- `write_efficiency =`
  - `#hits / #writes`

Исходные признаки:

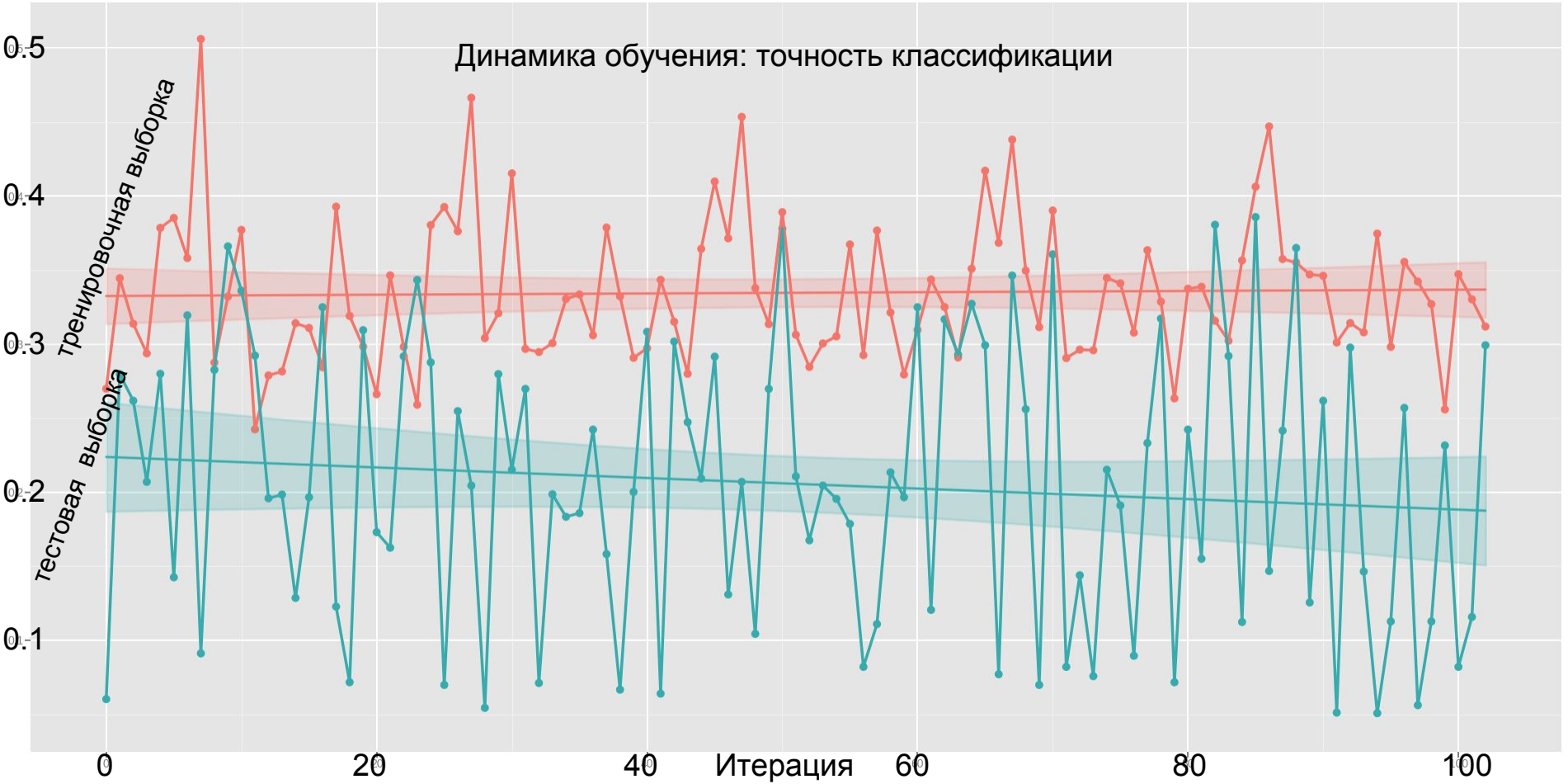
- Логический адрес первого блока (Logical Block Address)
- Число запрошенных блоков
- Тип запроса - запись или чтение
- Тип запроса - последовательный или случайный
- Время поступления запроса

```
lba           len r S timestamp
119755391744 257 w S 1463739509.336498
114460470692 112 r S 1463739509.337114
114398528743  80 r S 1463739509.337629
114376849933  96 r S 1463739509.337751
114424647007  32 r S 1463739509.338286
114376850029  16 r S 1463739509.338364
114424647039 176 r S 1463739509.338711
119755392001 255 w S 1463739509.338835
119755392256 257 w S 1463739509.338846
340632980788   8 w S 1463739509.339016
...
```

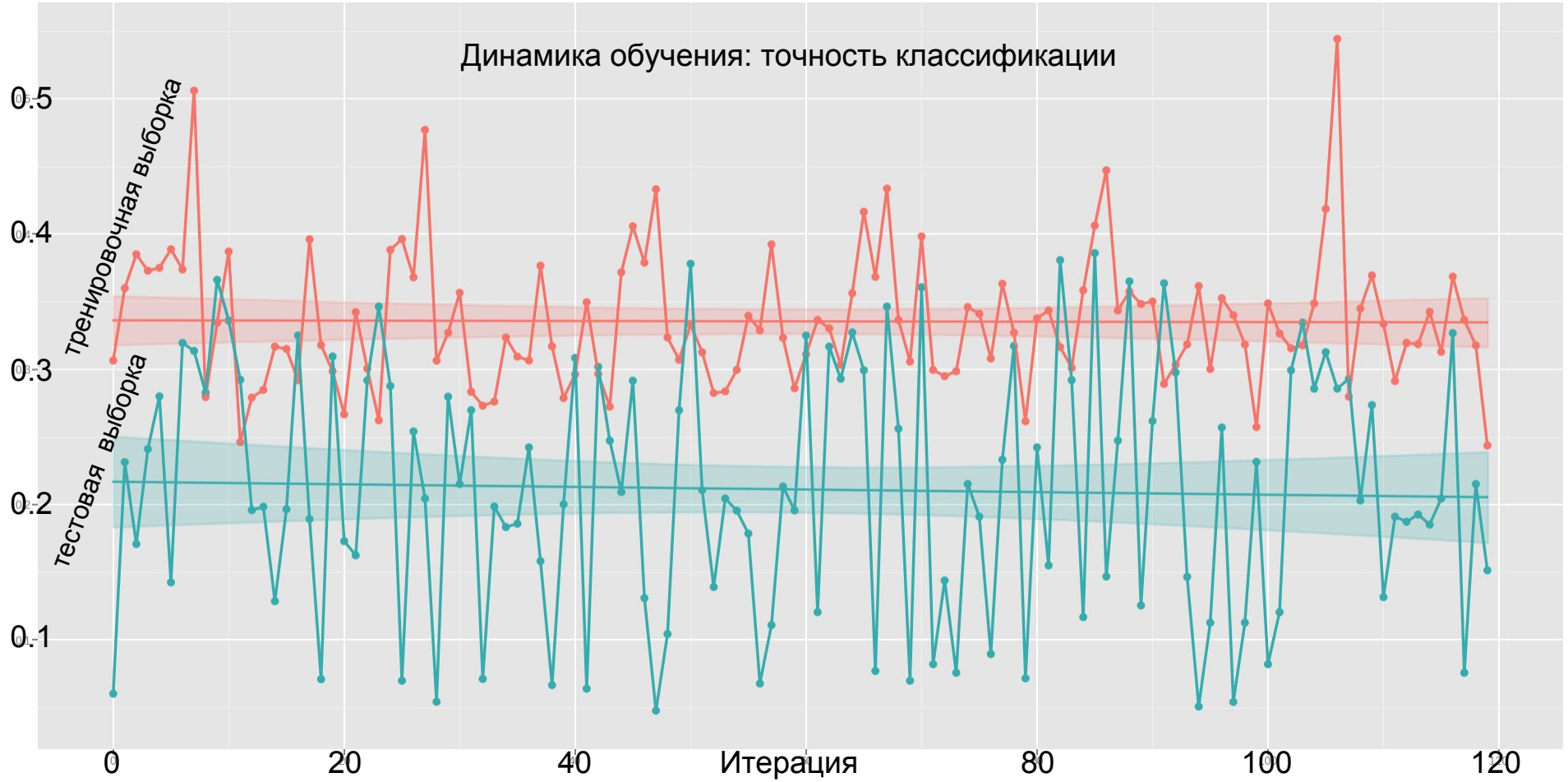
Эксперименты

1. Глубокое обучение
    a. Много сырых данных
    b. Пусть модели сами найдут представление данных
2. Деревья решений
    a. Разбиваем данные на большие блоки - мало данных
    b. Вычисляем на них признаки

Глубокое обучение: LSTM RNN



Глубокое обучение: Иерархическая нейронная сеть (HRNN)



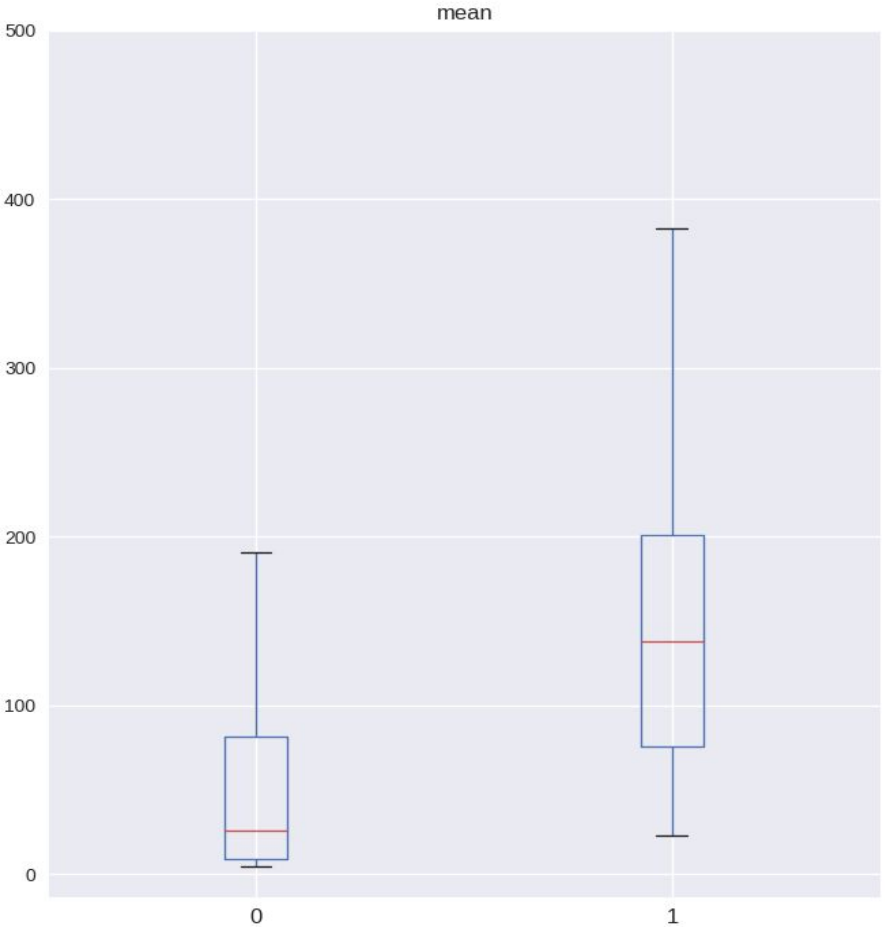
Эксперименты, деревья решений

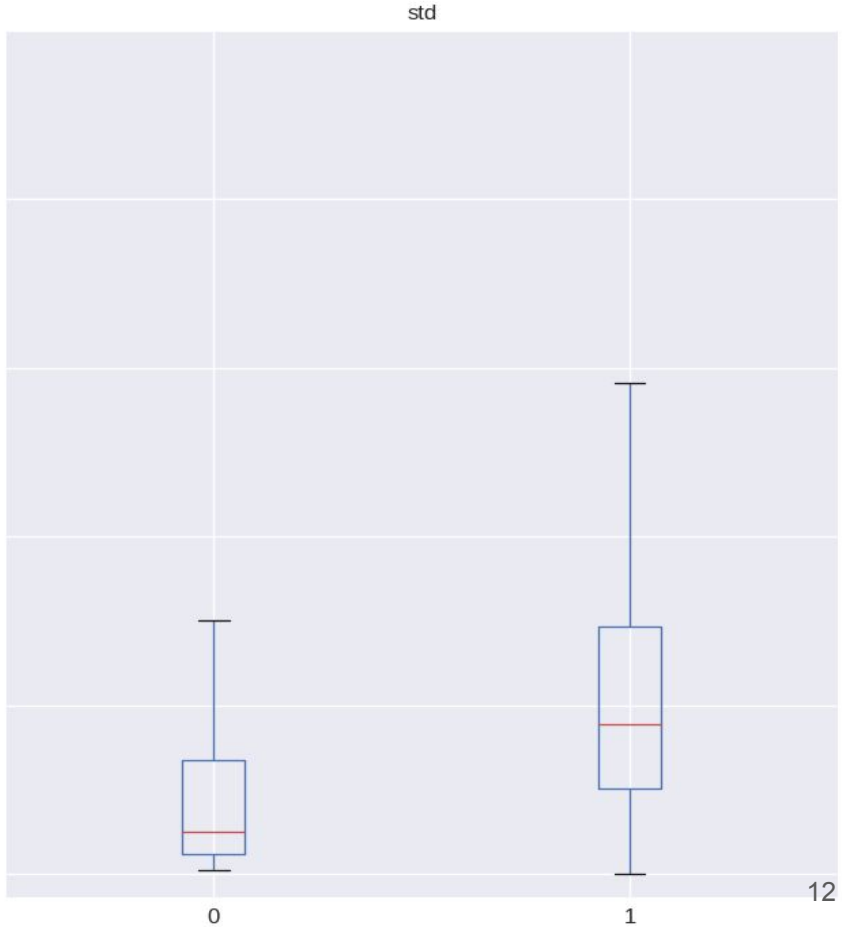
Представление данных (признаки):

- Частотные, разностные характеристики…
  - Подробно в тексте работы
- Много признаков с какими-то различиями между классами
  - Слабо коррелирующие признаки

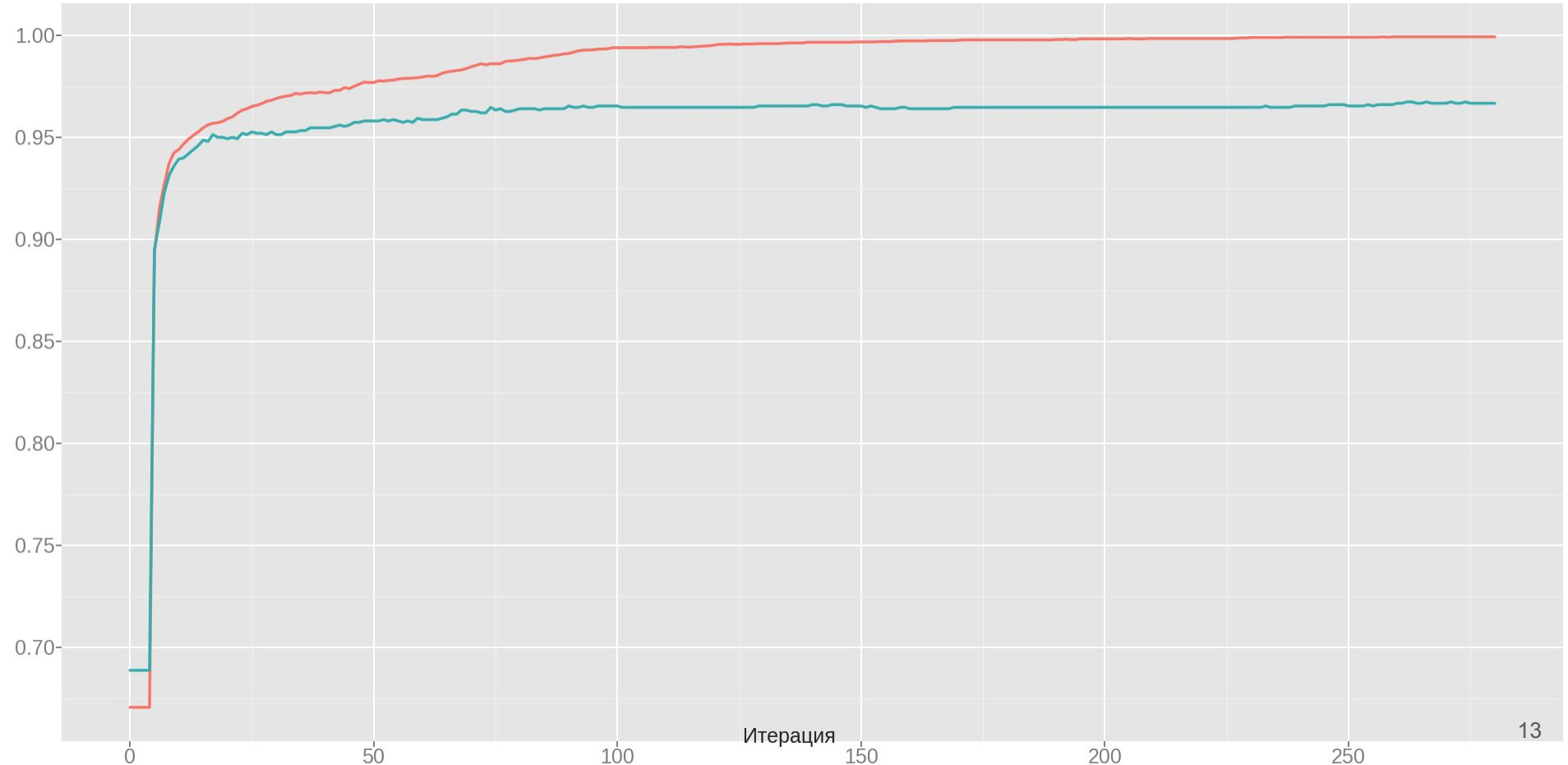
<moment>	count_lba	lba_interdiffs	count_difflba_1	count_len	len_interdiffs	count_difflen_1
kurtosis	3.798564	9.110780e-02	11.730286	18.752404	-0.904188	16.305275
mean	77.380000	2.524241e+11	172.400000	793.680000	205.731059	768.100000
skew	1.990029	-1.096458e+00	3.389365	4.448223	-0.107291	4.116541
std	76.304409	2.970855e+10	295.857660	2812.342146	5.967162	2486.047968

Различие между классами по признакам (пример)

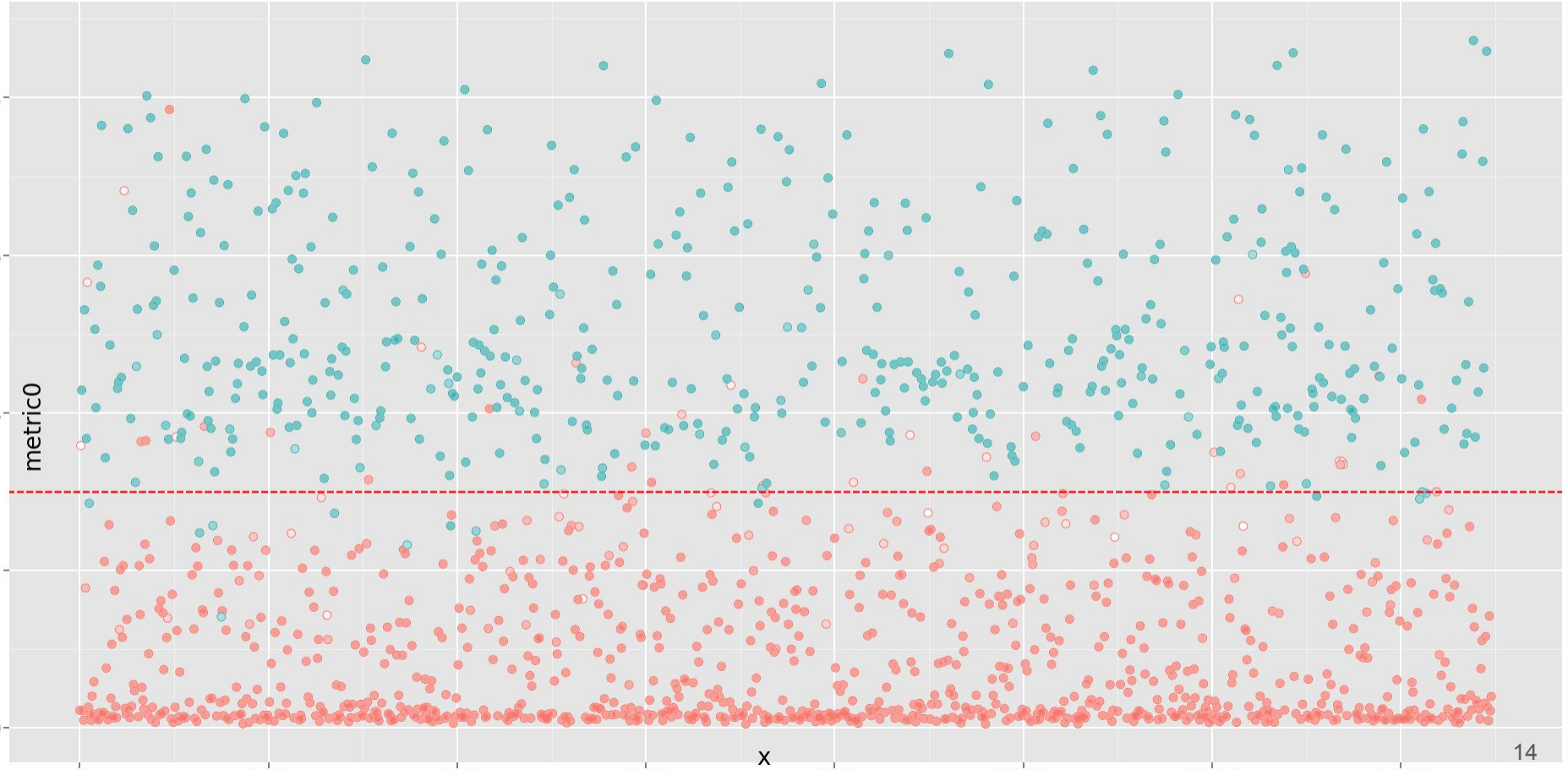




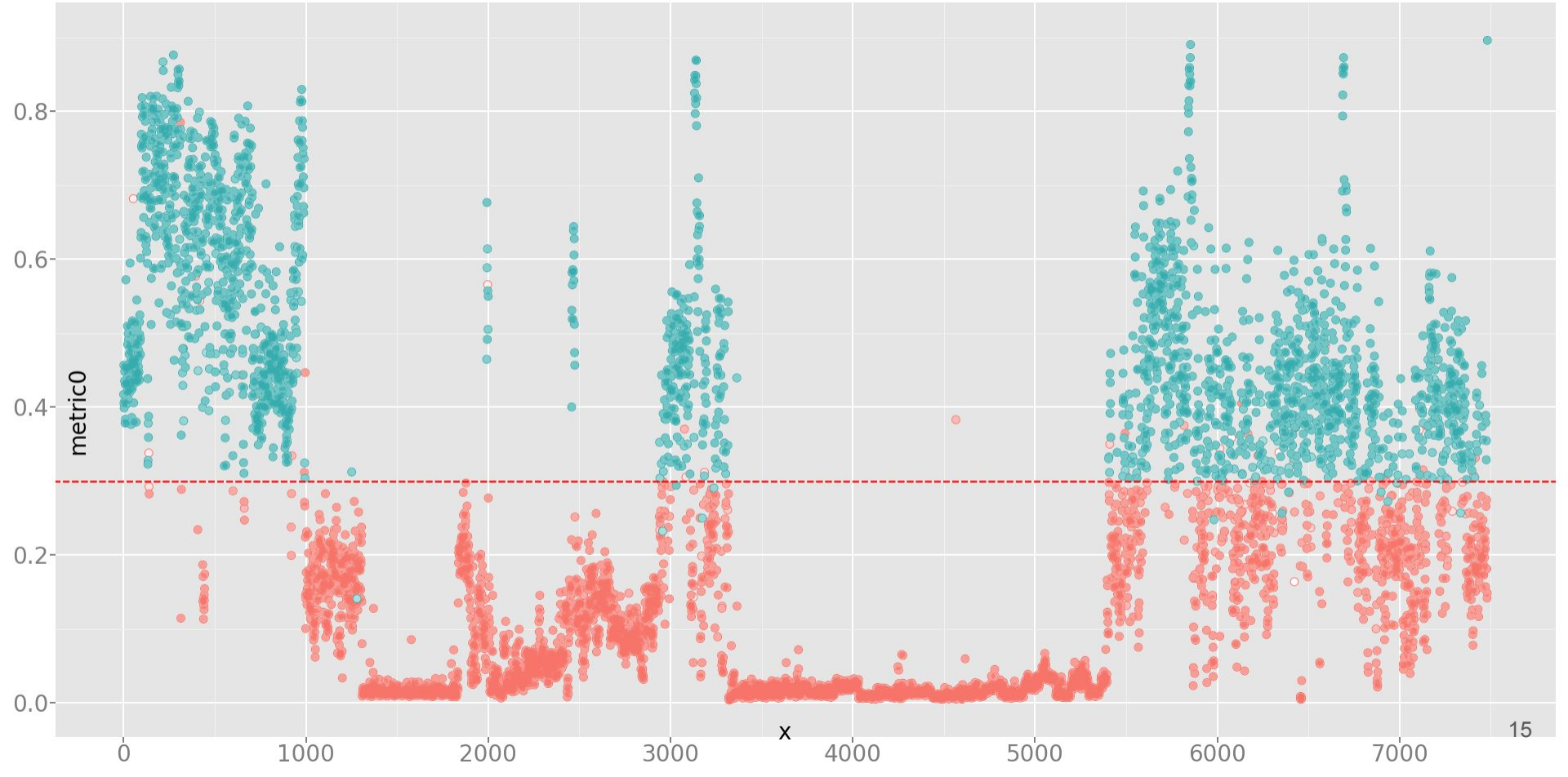
Динамика обучения: точность классификации



Прогноз на случайной тестовой выборке



Прогноз на всей выборке (тренировочная + тестовая)



Итоги экспериментов

- Представление данных — ключ к решению задачи
- Методы глубокого обучения не смогли его найти
- Задача решается стандартными методами с применением сложных признаков

Заключение

Выполнены следующие задачи:

- Проведен обзор подходящих методов машинного обучения
- Составлен план экспериментов
- Подготовлены данные для их проведения
- Проведены эксперименты
  - Получен положительный результат