

Histogram-Based Outlier Score (HBOS)

Подготовили студентки группы Э-2109 СПбГЭУ

Жданова С.Н., Кухтинова В.Ю., Темченко Д.Т., Чернова А.А.



Маркус Гольдштейн и Андреас Денгел

2012 год

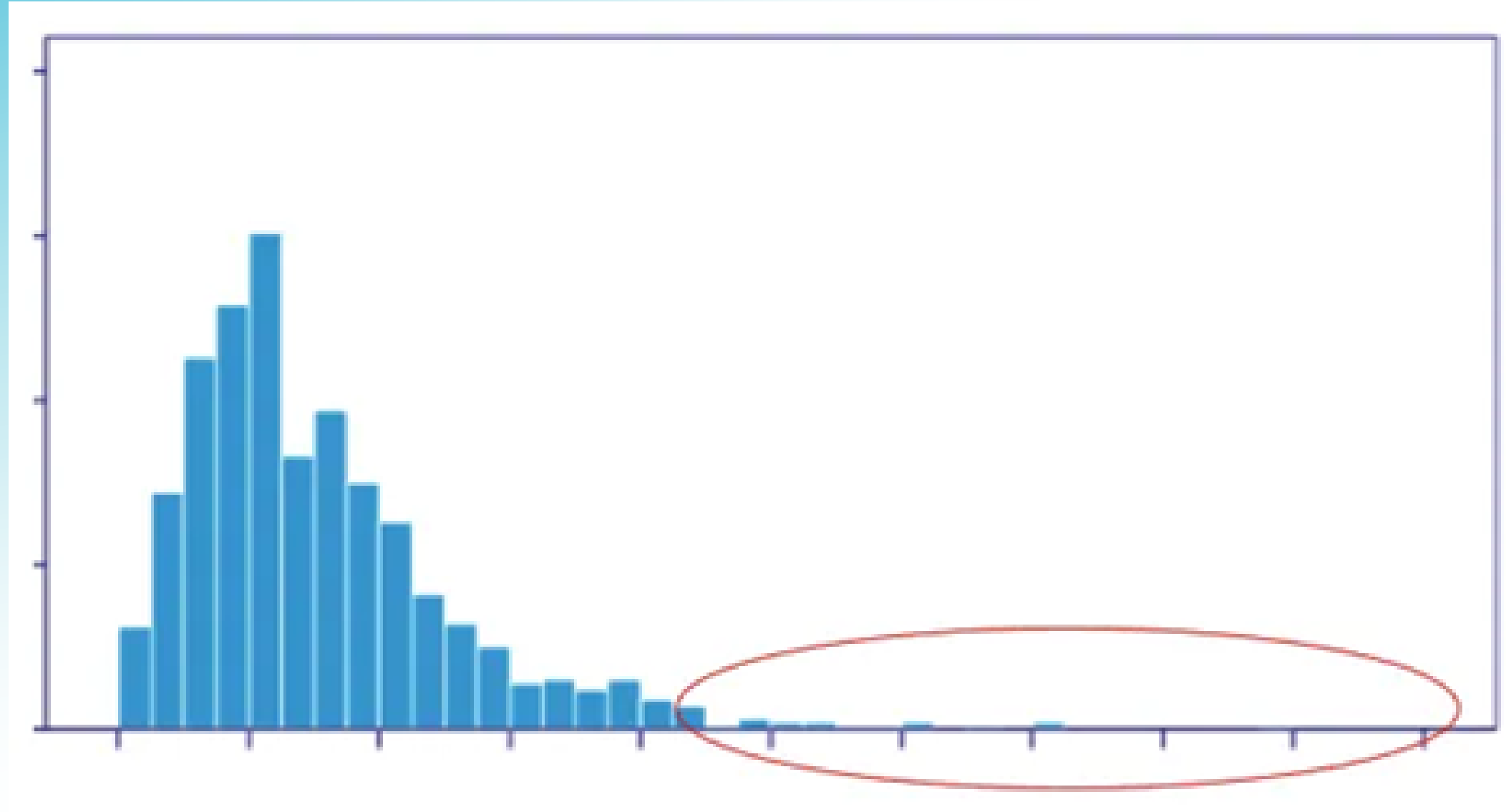
Способы обнаружения аномалий:

Контролируемое
Полуконтролируемое
Без учителя

Алгоритм:

1. Для каждого признака строится одномерная гистограмма.
2. Нормализация признаков.
3. Для каждого объекта p выборки рассчитывается HBOS

$$HBOS(p) = \sum_{i=0}^d \log \left(\frac{1}{hist_i(p)} \right) .$$



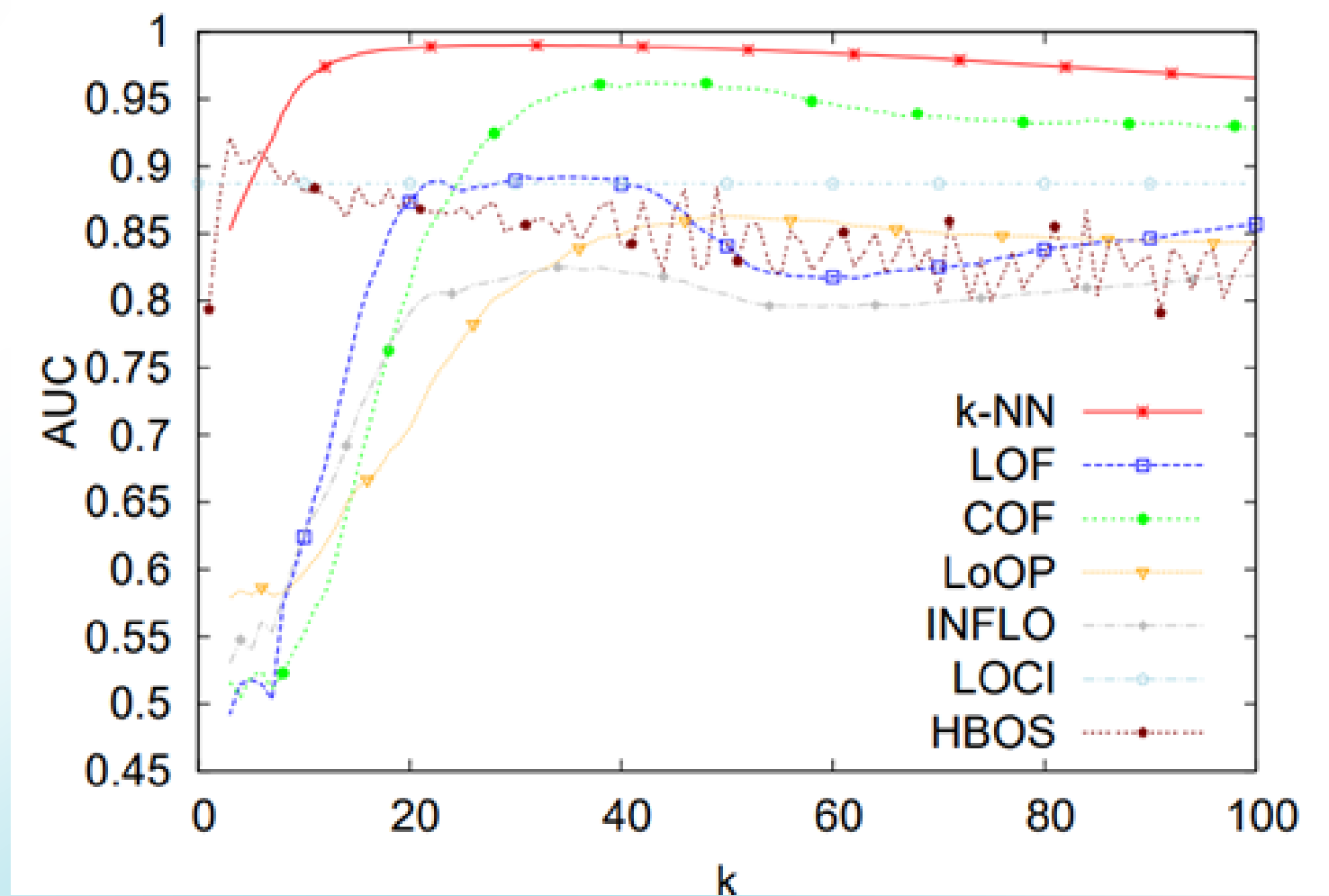
Точки высоким HBOS попадают в бины с низкими частотами -> значения встречаются редко

Преимущества и недостатки метода

- + легко реализовать и интерпретировать
- +эффективно для больших наборов данных
- +подходит для числовых данных
- может быть чувствителен к выбору количества бинов
- может не обнаруживать аномалии, которые имеют аномальные значения по отдельным признакам
- не подходит для категориальных и текстовых данных

Сравнение с другими методами

Algorithm	Breast-cancer	Pen-global	Pen-local
HBOS	0.9910	0.9214	0.7651
k -NN	0.9826	0.9892	0.9852
LOF	0.9916	0.8864	0.9878
Fast-LOF	0.9882	0.9050	0.9937
COF	0.9888	0.9586	0.9688
INFLO	0.9922	0.8213	0.9875
LoOP	0.9882	0.8492	0.9864
LOCI	0.9678	0.8868	—
CBLOF	0.8389	0.6808	0.7007
u-CBLOF	0.9743	0.9923	0.9767
LDCOF	0.9804	0.9897	0.9617



Сферы применения

- Мониторинг мошенничества
- Обнаружение аномалий в сетевом трафике
- Медицинская диагностика
- Контроль качества

Спасибо за внимание!