УДК 00.004.896

# СОПОСТАВЛЕНИЕ ОЦЕНЩИКОВ И АЛГОРИТМОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ ОБНАРУЖЕНИЯ АНОМАЛИЙ В НАБОРАХ ДАННЫХ

Пылов П. А., студент группы ИТб-162, IV курс
Протодьяконов А.В., [1] к.т.н., доцент
Кузбасский государственный технический университет имени Т.Ф. Горбачева
г. Кемерово

*В представленной работе охарактеризованы виды алгоритмов выявления выбросов и аномалий в пространственных наборах данных. Наборы данных содержат первый и второй режимы описания (именуемые областями высоких плотностей), позволяющие отобразить способности алгоритмов справляться с мультимодальными данными.*

Из датасета 15% всех данных заполнены случайным равномерным шумом. Эта пропорция является значением, заданным параметру в модели машинного обучения опорных векторов (OneClassSVM) и параметру загрязнения для других алгоритмов искусственного интеллекта и обнаружения выбросов. Границы принятия решения между аномалиями (выбросами) и сами выбросы маркированы черным цветом, кроме значения локального коэффициента выбросов, поскольку для него не создается метода прогнозирования, который будет применяться к новым данным, в то время, когда он используется для обнаружения выбросов.

Алгоритм опорных векторов (AOB) чувствителен к выбросам и поэтому не очень хорошо подходит для обнаружения выбросов. Этот алгоритм высокопроизводителен для определения новых данных (новые неизвестные числа), когда тренировочный набор не загрязнен выбросами. Тем не менее, обнаружение выбросов в высоких измерениях (более, чем 2) или без каких-либо знаний о распределении исходных данных является очень сложной задачей, и алгоритм опорных векторов с одним классом может выявить полезные результаты в этих ситуациях, в зависимости от значения его гиперпараметров (второй столбик сформированной матрицы рисунка 1)

Алгоритм сильной ковариации предполагает, что данные распределены по закону Гаусса и отображают эллипс. Таким образом, модель ухудшается, когда данные не являются одномодальными. В тоже время эта оценка устойчива к выбросам (первый столбик сформированной матрицы рисунка 1)

Алгоритм леса (Изолированный АЛ) и алгоритм ближайшего соседа очень производительно выполняют операции по мультимодальным датасе-

[1] Научный руководитель

там. Преимущество перед другими оценщиками показано для третьего набора данных, где два режима имеют разную плотность. Это преимущество объясняется локальным аспектом, означающим, что он сравнивает только оценку отклонения в одной выборке с оценками его соседей (третий столбик сформированной матрицы рисунка 1)

Наконец, для последнего набора данных (четвертый столбец рисунка 1) трудно сказать, что один образец более отклоняется от нормы, чем другой образец, поскольку они равномерно распределены в гиперкубе. За исключением модели опорных векторов, все оценщики представляют достойные решения для этой ситуации. В таком случае целесообразно более внимательно изучить оценки отклонений в выборках, ведь результат точного алгоритма оценки не зависит от данных (рисунок 1)

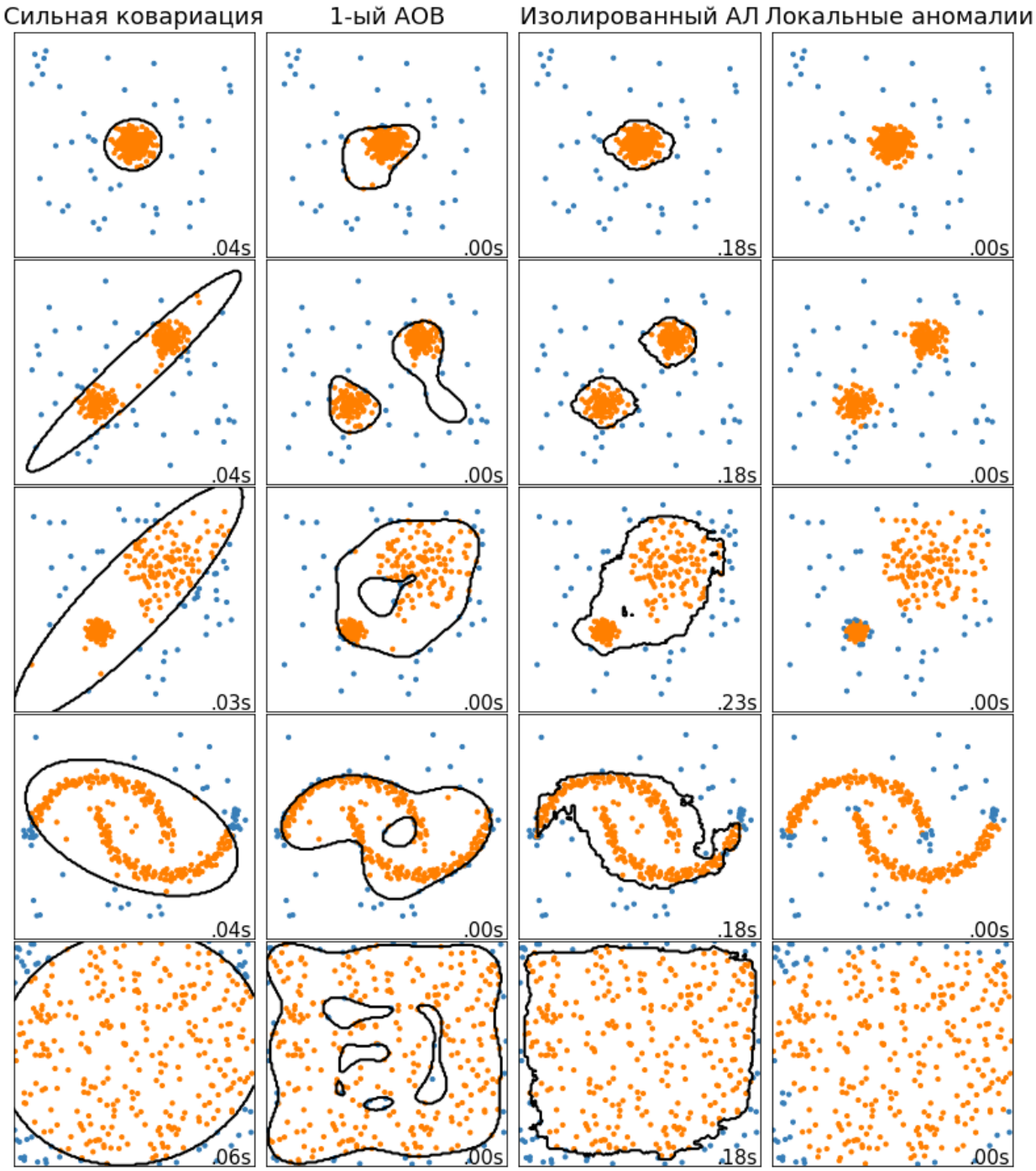


Рисунок 1. Матрица оценщиков алгоритмов искусственного интеллекта.

Изложенные в работе примеры дают некоторое понимание внутреннего устройства алгоритмов, но это понимание может отличаться от действительного на очень больших данных.

Наконец, параметры моделей здесь были выбраны вручную, но на практике их необходимо корректировать. При отсутствии помеченных данных проблема полностью не контролируется разработчиками, поэтому выбор модели может стать проблемой при условии отсутствия меток.

## Список литературы:

1. M. Narasimha Murty, V. Susheela Devy. Introduction to pattern recognition and machine learning, IISc press, New Jersey – London, 2015
2. S. Theodoridis, A. Pikrakis. Introduction to Pattern Recognition: A MATLAB Approach, Elsevier press, 2010
3. Christopher Bishop. Pattern Recognition and Machine Learning, Plenum press, New York – London, 1971
4. A. Geron. Hands-on Machine learning with Scikit-Learn, Keras, and Tensor-Flow, OREILLY Sebastopol, California – USA, 2019
5. P. Bruce, A. Bruce. Practical statistics for data scientists, OREILLY Sebastopol, California – USA, 2017