ПОИСК АНОМАЛЬНЫХ ЗНАЧЕНИЙ В MATLAB

Юхновская О.В.,
молодой ученый, БНТУ, г. Минск, Республика Беларусь
Научный руководитель – ***Гундина М.А.,*** *канд. физ.-мат. наук, доцент*

Ключевые слова. Аномальные значения, выборка, MatLab, статистическая совокупность.

Keywords. Anomalous values, sampling, MatLab, statistical universe.

В современных компьютерных системах появилась возможность реализовать автоматические пользовательские алгоритмы обнаружения аномальных значений выборки данных, полученных промышленным оборудованием, это осуществляется с помощью поиска и анализа закономерностей исходных эмпирических данных.

На данный момент существует множество подходов для определения аномалий и подходов к автоматизации процесса их выявления. Одни требуют привлечения сложного математического аппарата, другие достаточно непросты в компьютерной реализации, третьи должны учитывать особенности природы исходного анализируемого процесса.

Целью данного исследования является применение компьютерной системе MatLab для выявления аномальных значений выборки.

Процессом поиска аномальных значений несколько десятилетий интересуются физики, математики, астрономы, медики и другие ученые. Различные подходы широко представлены в отечественной и зарубежной литературе. Так, например, для поиска аномальных значений выборки могут быть использованы принципы классификации на основе репликационных нейронных сетей [1]. В основе такого подхода лежат принципы машинного обучения.

Для обнаружения аномалий выборки также может использоваться классификация на основе нейронных сетей глубинного обучения [2]. Этот подход использует глубокие нейронные сети для автоматического изучения выборки. Он также может быть использован для прогнозирования оценок аномалий каждого нового значения выборки.

Причина возникновения аномальных данных может быть связана с ошибками датчиков или особенностями процесса передачи этой информации. В этом случае могут быть использованы автоматизированные методы обнаружения аномалий с использованием байесовских сетей, которые выполняют быструю пошаговую оценку данных по мере их поступления, масштабируются до больших объемов данных и не требуют априорной информации о переменных процесса или типах аномалии [3]. Эти методы нашли свое широкое применение при анализе метеорологических данных [4]. Они позволят выявить аномалии выборки, которые вызваны двумя реальными событиями: отказом используемого датчика и сильным переменами в погодных условиях, например, штормом.

Для анализа аномалий в процессах функционирования предприятий может быть использована так называемая «классификация на основе правил» [5]. Постоянно меняющийся сетевой трафик выявляет новые типы подозрительной активности, которые могут быть небезопасны для ресурсов предприятия.

Существуют также алгоритмы, определяющие аномальные явления, на основе систем нечеткой логики. Нечеткая логика основана на обобщении классической логики и теории нечетких множеств для формализации нечетких знаний, характеризуемых лингвистической неопределенностью. В настоящее время такая логика широко применяется в вычислительных и информационных системах различного назначения. Она незаменима в тех случаях, когда на поставленные вопросы невозможно получить четкие ответы или заранее неизвестны все возможные ситуации. Такие алгоритмы представляют собой построенный автомат с конечным числом состояний, а два других отслеживают статистические отклонения от нормального поведения программы. Производительность этих алгоритмов оценивается в зависимости от количества доступных обучающих данных и сравнивается с хорошо известным методом обнаружения аномальных значений [6].

Понятие аномалии широко используется и в медицине. Врожденные аномалии представляют огромную опасность для больных и восприимчивых лиц и значительно ухудшают их жизнь. Они также широко исследуются, и анализируются пути их минимизации [7].

Существующие алгоритмы обнаружения аномалий на основе кластеризации. Они обучаются на немаркированных данных для обнаружения новых аномалий [8]. Кластерный анализ представляет собой совокупность различных алгоритмов деления объектов на группы, схожие по одному или нескольким признакам.

Отдельной группой выявления аномальных значений являются статистические методы. К ним относятся, например, параметрические методы. Они позволяют оценивать параметры модели в режиме реального времени, что устраняет необходимость в длительной фазе обучения или ручной настройке параметров [9].

Существует также подход к обнаружению аномалий на основе признаков [10], согласно которому строится гистограмма различных характеристик выборки, моделируется шаблон гистограмм, и выявляются отклонения от созданных моделей.

Материал и методы. В качестве компьютерной системы для организации процессы автоматизации процесса определения аномальных значений выбрана система MatLab. Она позволяет применять методы машинного обучения решения данной задачи.

Результаты и их обсуждение. Загрузим исходный файл file1.mat.

Для этого используем встроенную функцию load. Исходный файл содержит тренировочную выборку data и тестируемые данные datatest. Предполагаем, что выборке data аномальных значений нет. Осуществляем обучение с помощью метода изолирующего леса:

```
rng(“default”)
[Mdl,tf,s]=iforest(data);
```

Функция iforest возвращает индикаторы аномалий tf.

Для поиска аномалий для datatest используем команду:

```
[tf_test,s_test]=isanomaly(Mdl,datatest);
```

Для построения гистограммы используется команда:

```
histogram(s,Normalization="probability")
hold on
histogram(s_test,Normalization="probability")
xline(Mdl.ScoreThreshold,"r-",join(["Threshold" Mdl.ScoreThreshold]))
legend("Training Data","Test Data",Location="northwest")
hold off
```

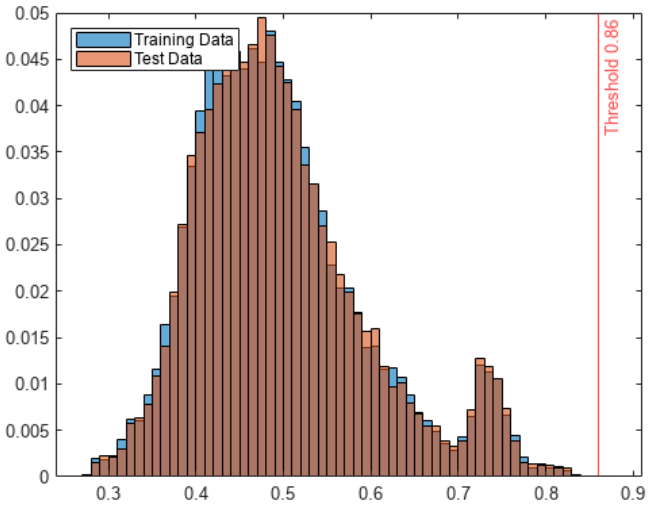

Рисунок 1 – Исходные и аномальные значения

Заключение. Процесс выявления аномалий является очень важным вопросов в задачах предсказания поломок оборудования, выявления аномального спроса на потребляемую продукцию, выявления нестандартного поведения информационно-измерительной системы. В компьютерной системе Mathematica имеются встроенные функции анализа аномальных значений выборки. Также эта система позволяет обрабатывать массивы большого объема.

Выбор метода для детектирования аномалий зависит в первую очередь от поставленной задачи, данных и имеющейся априорной информации. Рассмотренные подходы являются лишь математическими моделями понятия аномальности и отталкиваются от интерпретации задачи.

Самым стабильным и общим алгоритмом остаётся Isolation Forest, который не требует никакой априорной информации. При этом какие-либо модификации, «метрически исправляющие» пространство признаков, для него не нужны и чаще приводят к потере качества.

1. Гундина, М. А. Выявление аномальных кластер выборки в компьютерной системе Wolfram Mathematica / М. А. Гундина // Вестник Белорусско-Российского университета. – 2022. – №. 4 (77). – С. 75–83.

2. Ghosh, S. Network anomaly detection using a fuzzy rule-based classifier / S. Ghosh, A. Pal, A. Nag, S. Sadhu, R. Pati // Computer, Communication and Electrical Technology. – 2017. – P. 61–65.

3. Hill, D. J. Real-Time Bayesian Anomaly Detection in Streaming Environmental Data / D. J. Hill, B. S. Minsker, E. Amir // Water resources research. – 2009. DOI: 10.1029/2008WR006956.

4. Звягин, Л.С. Байесовский подход в современном экономическом анализе и имитационном моделировании / Л. С. Звягин // Мягкие измерения и вычисления. – 2018. – No.1. – С. 17–26.

5. Nasr, A. An Intrusion Detection and Prevention System based on Automatic Learning of Traffic Anomalies / A. Nasr, M. Ezz, M. Abdulmaged // International Journal of Computer Network and Information Security. – 2016. – No. 8. – P. 53 –60.

6. Michael, C. Simple, State-Based Approaches to Program-Based Anomaly Detection // C. Michael, A. Ghosh // ACM Trans. Inf. Syst. Secur. – 2002. – No. 5. – P. 203–237.

7. Amer, A. An overview on congenital dental anomalies / A. Amer [and etc.] // International Journal of Community Medicine And Public Health. – 2002. – V.9. –No. 2. – P. 976–980.

8. Portnoy, L. Intrusion Detection with Unlabeled Data Using Clustering / L. Portnoy, E. Eskin, S. J. Stolfo. – New York: Columbia University, 2001. – 38 p.

9. Thatte, G. Parametric Methods for Anomaly Detection in Aggregate Traffic / G. Thatte, U. Mitra, J. Heidemann // ACM Transactions on Networking. – 2011. – V. 19(2). – P. 512–525.

10. Kind, A. Histogram-based traffic anomaly detection / A. Kind, M.P. Stoecklin, X. Dimitropoulos // IEEE Transactions on Network and Service Management. – 2009. – V. 6(2). – P. 110–121.