3. Гороховатский, В. А. Структурный анализ и интеллектуальная обработка данных в компьютерном зрении / В. А. Гороховатский. – Харьков: Компания СМИТ, 2014. – 316 с.

4. Жданович, М. Н. Обработка изображений. Расстояние Махаланобиса / М. Н. Жданович, М. А. Гундина // Новые направления развития приборостроения : материалы 15-й Междунар. научн.-технич. конфер. мол. учен. и студ., Минск, 20−22 апреля 2022 г. / Белорусский национальный технический университет ; редкол.: О. К. Гусев (пред. редкол.) [и др.]. – Минск : БНТУ, 2022. – С. 205–206.

УДК 51

РЕАЛИЗАЦИЯ СТАТИСТИЧЕСКИХ АЛГОРИТМОВ ОПРЕДЕЛЕНИЯ АНОМАЛЬНЫХ ЗНАЧЕНИЙ ВЫБОРКИ В WOLFRAM MATHEMATICA

Гундина М.А., Кондратьева Н.А., Каменко Д.А.

Белорусский национальный технический университет
Минск, Республика Беларусь

Аннотация. В данной статье рассматривается реализация некоторых статистических алгоритмов выявления аномальных значений выборки, реализованных в компьютерной системе WolframMathematica.
Ключевые слова: аномальное значение, выборка, компьютерная система Wolfram Mathematica.

IMPLEMENTATION OF STATISTICAL ALGORITHMS FOR DETERMINATION ANOMALOUS SAMPLE VALUES IN WOLFRAM MATHEMATICA

Hundzina M., Kondratyeva N., Kamenko D.

Belarusian National Technical University
Minsk, Republic of Belarus

Abstract. This article discusses the implementation of some statistical algorithms for detecting anomalous sample values, implemented in a computer system Wolfram Mathematica.
Key words: anomalous value, sample, computer system.

Адрес для переписки: Гундина М.А., пр. Независимости, 65, Минск 220013, Республика Беларусь
e-mail: hundzina@bntu.by

Задача автоматизации процесса выявления аномальных значений выборки решалась в инженерной практике и математической статистике и не теряет свою актуальность несколько десятилетий [1–4]. Известно, что аномальные значения способны существенно исказить функционирование математических моделей анализа данных, что может привести к снижению надежности и некорректной работе всей системы [5, 6].

Единицы статистической совокупности, у которых значения анализируемого признака существенно отклоняются от основного массива, называются аномальными явлениями, «грубыми ошибками» или выбросами. Возможны и аномальные результаты, обусловленные сбоями при измерениях и регистрации данных, резкими отклонениями условий наблюдений, нештатной работой оборудования, ошибками операторов и др. Наличие аномальных результатов может привести к недостоверным результатам при оценивании и контроле соответствия характеристик системы предъявляемым требованиям. Поэтому необходимо выявлять и устранять аномальные результаты измерений [7].

Аномальные значения определяются исходя из их характеристики по отношению ко всей совокупности. Эти значения можно разделить на три группы:

1) значения, отражающие объективное развитие процесса, но сильно отклоняющиеся от общей тенденции;

2) значения, возникающие вследствие изменения методики расчетов;

3) значения, возникающие из-за ошибок при измерении показателя, при записи, передаче информации и т.д.

Поэтому процесс выявления и затем удаления этих значений состоит из нескольких этапов [8]. Вначале выявляются значения, которые выходят за границы интервала возможного варьирования характеристики признака.

Исходя из физического смысла исследуемой величины, могут рассматриваться как аномальные значения те, которые не соответствуют монотонному характеру изменения величины при последовательных наблюдениях, а также значения, приращения которых превышают предельно возможную скорость изменения величины.

Одной из групп методов выявления аномальных значений является группа статистических методов [2, 4].

Реализация метода поиска аномалий с помощью правила сигм. С помощью этого метода можно осуществлять контроль за нахождением параметра в допустимых границах, что удобно в производственных процессах. Анализ выбросов в

данных позволяет определить аномальные значения в нестационарных рядах с распределением близким к нормальному закону распределения. Основу данного метода анализа составляет расчет среднего значения ряда и среднеквадратичного отклонения.

Формула для вычисления среднего значения ряда:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \,, \qquad (1)$$

где n – количество элементов выборки, x_i – i-й элемент выборки.

Формула для вычисления среднеквадратичного отклонения:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \,. \qquad (2)$$

Суть данного метода сводится к тому, что любые значения ряда, отличающиеся от среднего больше, чем на два σ, являются потенциальными аномалиями. Порог определения аномалий задается формулой:

$$T = x_i \pm 2\sigma \,. \qquad (3)$$

Алгоритм. Строим гистограмму. Из графика распределения видно, что в исходных данных присутствуют значения явно отстоящее от остальных.

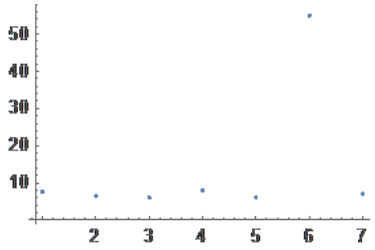

Рисунок 1 – Значения исходной выборки

Определяем пороги T_1 и T_2 следующим образом:

```
T1=Mean[v[[All,2]]]-2 StandardDeviation[v[[All,2]]];
T2=Mean[v[[All,2]]]+2 StandardDeviation[v[[All,2]]]
```

Эти значения являются порогами классификации значения как аномалии. Замети, что если аномалий оказалось много, имеет смысл увеличить порог – задать его равным 3×σ, 4×σ3×σ, 4×σ и более.

В цикле проверяем весь массив значений на наличие аномальных и заносим эти значения в новый массив *s*:

```
For[s={};i=1,i<=Length[v],i++,If[Or[v[[i,2]]<=
T1,v[[i,2]]>=T2],s=Append[s,v[[i]]] ]].
```

Кроме метода сигм простым методом обнаружения грубых ошибок считается метод, на основании Т – Критерия Граббса. Критерий Граббса определен для следующих гипотез:

H0: В наборе данных нет выбросов.

H1: В наборе данных присутствует как минимум один выброс.

$$T_k = \frac{|x - \bar{x}|}{s} \,. \qquad (4)$$

Значение критерия Граббса показывает максимальное абсолютное отклонение от выборочного среднего в единицах среднеквадратичного отклонения.

```
For[s1={};i=1,i<=Length[v],i++,If[Abs[v[[i,2]]-
Mean[v[[All,2]]]]/(StandardDeviation
[v[[All,2]]]/Length[v]-1)==Max[s],s1=
Append[s1,v[[i]]]]]
```

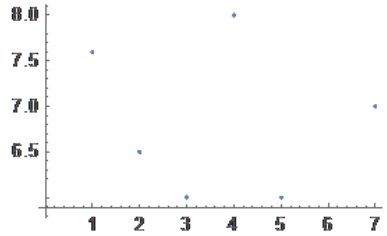

Рисунок 2 – Результат удаления значения из массива

Данный критерий можно использовать для выделения аномальных результатов измерений только в случае нормального закона.

Критерий Граббса определяет один выброс за одну итерацию. Этот выброс исключается из набора данных и тест повторяется до тех пор, пока не будут обнаружены все выбросы.

Литература

1. Вероятность и математическая статистика: Энциклопедия / под ред. Ю. В. Прохорова. М.: Большая Российская энциклопедия, 2003. – 911 с.

2. Линник, Ю. В. Метод наименьших квадратов и основы математико-статистической теории обработки наблюдений / Ю. В. Линник. – М.: Физматгиз, 1962. – 352 с.

3. Новицкий, П. В. Оценка погрешностей результатов измерений / П. В. Новицкий, И. А. Зограф. – Ленинград: Энергоатомиздат, 1985. – 248 с.

4. Смирнов, Н. В. Курс теории вероятностей и математической статистики для технических приложений / Н. В. Смирнов, Дунин-Барковский В. М. – Наука, 1965. – 552 с.

5. Богатырев, В. А. Надежность мультикластерных систем с перераспределением потоков запросов / В. А. Богатырев, С. В. Богатырев // Известия высших учебных заведений. Приборостроение. – 2017. – Т. 60, № 2. – С. 171–177.

6. Контроль и безопасность функционирования дублированных компьютерных систем / В. А. Богатырев [и др.] // Научно-технический вестник информационных технологий, механики и оптики. – 2017. – Т. 17, № 2. – С. 368–372.

7. Лукин, В. Л. Статистический метод выявления аномальных результатов измерений характеристик технических систем / В. Л. Лукин, Б. И. Сухорученков, В. И. Кузнецов // Двойные технологии. – 2010. – № 2 (51). – С. 32–40.

8. Сухорученков, Б. И. Методы анализа характеристик летательных аппаратов / Б. И. Сухорученков, В. А. Меньшиков. – М.: Машиностроение, 1995. – 475 с.