

ОСОБЕННОСТИ ОПРЕДЕЛЕНИЯ АНОМАЛЬНЫХ ЗНАЧЕНИЙ В СИСТЕМЕ WOLFRAM MATHEMATICA

Д.А. Каменко¹, М.А. Гундина², М.Н. Жданович³

¹Белорусский национальный технический университет,
пр. Независимости, 65, 220013, г. Минск, Беларусь, dimakamenko.2000@gmail.com.

²Белорусский национальный технический университет,
пр. Независимости, 65, 220013, г. Минск, Беларусь, hundzina@bntu.by

³«Отраслевая лаборатория новых технологий и материалов»,
ОАО "ИНТЕГРАЛ" – управляющая компания холдинга "ИНТЕГРАЛ",
ул. Казинца, 121А, 220108, г. Минск, Беларусь, MZhhdanovich@integral.by

В статье приводится обзор существующих методов определения аномальных значений выборки. Рассматривается автоматизация метода, основанного на расстоянии Махаланобиса, в компьютерной системе Wolfram Mathematica. Эта система позволяет автоматизировать процесс поиска аномальных значений выборки. Также приводится пример использования сингулярных матриц для выявления аномальных значений в выборках, представленных в виде числовых матриц.

Ключевые слова: Аномальные значения; сингулярная матрица; расстояние Махаланобиса; Wolfram Mathematica.

FEATURES OF DEFINITION OF ANOMALOUS VALUES IN WOLFRAM MATHEMATICA SYSTEM

D.A. Kamenka^a, M.A. Hundzina^b, M.N. Zhdanovich^c

^a Belarusian National Technical University, 65 Nezavisimosti Ave., Minsk 220013,
Belarus, dimakamenko.2000@gmail.com.

^b Belarusian National Technical University, 65 Nezavisimosti Ave., Minsk 220013,
Belarus, hundzina@bntu.by

^c «Branch laboratory of technologies and materials», JSC "INTEGRAL" – the management
company of the holding "INTEGRAL",
121a St.Kazintsa, Minsk 220108, Belarus, MZhhdanovich@integral.by

The article provides an overview of existing methods for determining the anomalous values of the sample. The automation of the method based on the Mahalanobis distance in the Wolfram Mathematica computer system is considered. This system allows you to automate the process of searching for anomalous sample values. An example of using singular matrices to detect outliers in samples presented as numerical matrices is also given.

Keywords: Anomalous values; singular matrix; Mahalanobis distance; Wolfram Mathematica.

Введение

Задача автоматизации процесса выявления аномальных значений выборки не теряет свою актуальность несколько десятилетий и находит свое применение в экономике, инженерии и других отраслях науки и техники [1–4].

Аномальные значения способны существенно исказить функционирование математических моделей анализа данных, что может привести к снижению надежности и некорректной работе всей системы, неточности прогнозов, которые будут делаться на основе таких моделей [5, 6]. Также наличие аномальных результатов может привести к недостоверным результатам при оценивании и контроле соответствия характеристик оборудования предъявляемым требованиям.

Под аномальными значениями будем понимать единицы статистической совокупности, у которых значения анализируемого признака существенно отклоняются от основного массива. Такие значения также называют выбросами.

Причины возникновения аномальных результатов разные. Такие результаты могут быть обусловлены сбоями при измерениях и регистрации данных, резкими отклонениями условий наблюдений, ошибками операторов. Поэтому необходимо выявлять и устранять аномальные результаты измерений [7].

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.

1. Теоретические основы

В системе *Wolfram Mathematica* есть несколько встроенных функций, которые позволяют анализировать наличие аномальных значений. Так, например, функция системы *FindAnomalies* позволяет найти члены выборки, которые считаются аномальными по отношению выборке. Задавая порог принятия можно в автоматическом режиме определять аномальные значения выборки.

Данная функция может быть использована для многих типов данных, включая числовые, строковые и графические данные.

В системе также есть и другие функции, позволяющие анализировать аномальные значения. Функция *DeleteAnomalies* выдает новый набор данных, в котором уже удалены аномальные значения. Функции *AnomalyDetection* и *AnomalyDetectorFunction* позволяют проверить, является ли новое предъявляемое значение аномальным.

Степень аномальности значения может также определяться по значению расстояния Махаланобиса. Эта величина в математической статистике является мерой расстояния между векторами случайных величин. Она обобщает понятие евклидова расстояния.

Примем в рассмотрение предположение о нормальном законе распределения исходной выборки.

Для векторизации вычисления расстояний Расстояние Махаланобиса между двумя точками – это мера расстояния между двумя случайными точками U и V , одна из которых может принадлежать некоторому классу с матрицей ковариации COV :

$$d_m(U, V, COV) = \sqrt{(U - V)COV^{-1}(U - V)^T}, \quad (1)$$

где символ T обозначает операцию транспонирования, а под COV^{-1} подразумевается матрица, обратная ковариационной матрице.

Элементы ковариационной матрицы вычисляются следующим образом:

$$cov_{a,b} = \frac{1}{|C| - 1} \sum_{x \in C} (X_1 - \mu_1)(X_2 - \mu_2), \quad (2)$$

где μ_1, μ_2 – математические ожидания по признакам, $|C|$ – количество точек в классе.

Расстояние Махаланобиса широко применяется в задачах кластеризации и классификации в задачах определения соответствия точки известному классу. Оно отличается от расстояния Евклида тем, что учитывает корреляции между переменными и инвариантно масштабу.

2. Результаты и их обсуждение

Метод, основанный на расстоянии Махаланобиса

Точка, имеющая наибольшее расстояние Махаланобиса до остального множества точек, считается аномалией. Такая точка имеет наибольшее влияние на кривизну и на коэффициенты уравнения регрессии. Также расстояние Махаланобиса может быть использовано в задаче определения многомерных выбросов.

Пусть исходная выборка имеет вид:

```
cohort={{5.04,14.22},{5.50,5.83},{5.19,4.61},{4.78,4.12},{5.08,5.99},{4.29,4.18},{5.08,6.90}};
```

```
\[Mu]=Mean@cohort;s=Covariance@cohort;
```

Функция ListPlot позволяет представить точки в декартовой системе координат, дополнительно указать значения расстояния Махаланобиса для каждой точки выборки:

```
ListPlot[{cohort,Labeled[#,Round[Dm[#,\[Mu],s],.01]]&/@points},PlotRange->All,AspectRatio->1,PlotStyle->{Darker@LightBlue,{Red,PointSize[.01]}}].
```

Результат работы алгоритма представлен на рис.1.

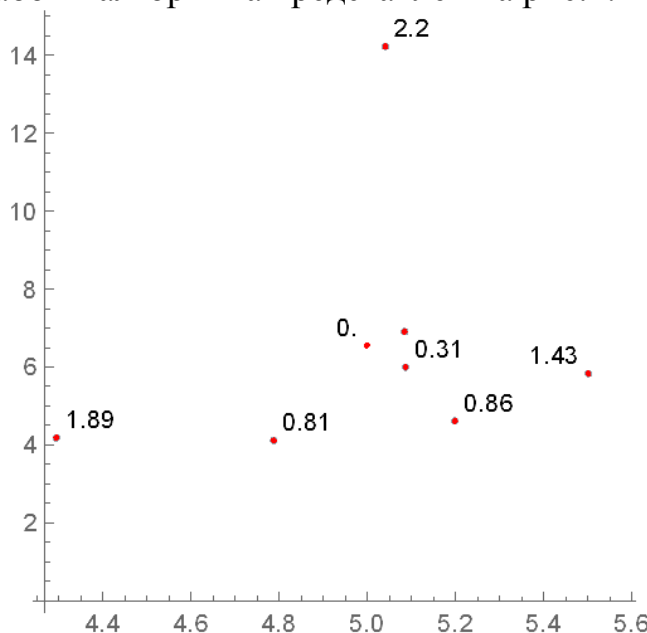


Рисунок 1 – Графическое представление исходной выборки

Анализируя полученные значения расстояний, было выявленное одно аномальное значение, величина расстояния для этого значения значительно превышает эти же расстояния для других значений выборки.

Метод, основанный на разложении матрицы

Основная идея метода обнаружения аномальных значений, основанного на разложении матриц, состоит в том, чтобы использовать сингулярное разложение исходной матрицы данных.

Наилучшая матрица получается из сингулярного разложения матрицы M по формуле:

$$M=ULV^T. \quad (3)$$

L – матрица размера $m \times n$ с неотрицательными элементами, у которой элементы, лежащие на главной диагонали – это сингулярные числа (а все элементы, не лежащие на главной диагонали, являются нулевыми), а матрицы U и V – это две унитарные матрицы, состоящие из левых и правых сингулярных векторов соответственно.

В системе Wolfram Mathematica сингулярное разложение может быть получено с помощью следующей формулы:

$$\{u, l, v\} = \text{SingularValueDecomposition}[M1];$$

Приближенная матрица $M_k = U_k L_k V_k^T$, U_k , L_k , V_k получаются из матриц сингулярного разложения отсечением до k первых столбцов.

Тогда приближенная матрица имеет вид, представленный на рис. 2.

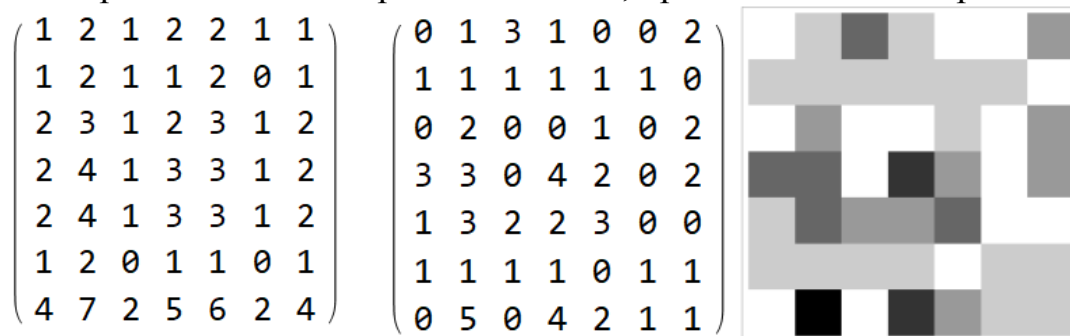


Рисунок 2 – Приближенная матрица. Матрица разности исходной матрицы с приближенной матрицей. Аномальные значения, выделенные темным цветом

Элементы, которые сильно отличаются от соответствующих элементов матрицы небольшого ранга, будут считать аномальными.

Заключение

В системе Wolfram Mathematica существуют как встроенные функции, позволяющие анализировать аномальные значения выборки, так и существует возможность автоматизировать процесс определения аномальных значений методами, основанными на расстоянии Махаланобиса и разложении матрицы.

Библиографические ссылки

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. М.: Физматгиз, 1962. 352 с.
3. Новицкий П.В., Зиграф И.А. Оценка погрешностей результатов измерений. Ленинград: Энергоатомиздат, 1985. 248 с.

4. Смирнов Н. В., Дунин-Барковский И. В. Курс теории вероятностей и математической статистики для технических приложений. М.: Наука, 1965. 552 с.
5. Богатырев В.А., Богатырев С.В. Надежность мультикластерных систем с перераспределением потоков запросов // Известия высших учебных заведений. Приборостроение. 2017. Т. 60. № 2. С. 171-177.
6. Богатырев В.А., Винокурова М.С., Петров П.А., Назарова М.Л., Шабakov Р.В. Контроль и безопасность функционирования дублированных компьютерных систем // Научно-технический вестник информационных технологий, механики и оптики. 2017. Т. 17. № 2. С. 368-372.
7. Лукин В.Л., Сухорученков Б.И., Кузнецов В.И. Статистический метод выявления аномальных результатов измерений характеристик технических систем // Двойные технологии. 2010. № 2(51). С. 32-40.
8. Сухорученков Б. И., Меньшиков В. А. Методы анализа характеристик летательных аппаратов. М.: Машиностроение, 1995. 475 с.