УДК 519.254

Сравнительный анализ методов интеллектуальной обработки данных для повышения качества прогнозных моделей

О. В. Рычка
Донецкий национальный технический университет, г. Донецк
E-mail: olga_rychka@mail.ru

Аннотация
В данной статье отмечена важность предварительной обработки в анализе данных. Описаны результаты сравнительного анализа эффективности различных методов поиска аномальных значений в статистических данных и предложенного автором метода. Представлена программная реализация предложенного метода. Реализованный в работе метод обнаружения и обработки выбросов позволяет определять более точные значения различных показателей и способствует построению достоверных прогнозов.

Введение

Важным этапом анализа данных является их предварительная обработка с целью идентификации значений, которые не соответствуют модели поведения анализируемого процесса. Такие значения называют аномалиями. Одним из значимых инструментов анализа данных является регрессионный анализ – статистический метод, позволяющий выявлять соотношения между зависимой переменной и одной или несколькими независимыми переменными [1].

Основными двумя направлениями поиска аномалий является обнаружение выбросов и обнаружение новизны. В отличие от выбросов, новизна указывает на определённые изменения в системе и не является следствием ошибок в данных. В этом случае, задача заключается в своевременном обнаружении аномалий и анализе причин их появления, поскольку они могут сигнализировать о критически важных событиях. Целью такого поиска может являться обнаружение неисправности функционирования оборудования, сетевых хакерских атак, мошенничества с банковскими картами, выявление изменений в показателях здоровья человека, и т.д. Т.е. в данном случае исследователя интересуют сами аномалии, как индикаторы отклонения от нормального поведения системы и причины их возникновения. Поэтому, после обнаружения такие данные подвергаются дальнейшему анализу.

Основные причины возникновения выбросов – неточные измерения, некорректный ввод данных, выход из строя оборудования и т.д. В этом случае, после обнаружения аномалий их следует подвергнуть дальнейшей обработке – исключить из выборки или откорректировать [2]. Это позволит построить адекватную модель, наиболее точно описывающую существующую зависимость.

Чтобы репрезентативность выборки не была снижена, исключение аномалий из неё можно осуществлять, когда она содержит достаточное количество данных.

Основными методами корректировки являются:

- ручная замена выброса на другое, более подходящее значение;
- изменение экстремальных значений на наиболее вероятное значение;
- сглаживание данных;
- интерполяция аномалий. Они заменяются значениями, которые получены на основе ближайших соседей.

Целью исследования является проведение сравнительного анализа эффективности различных методов поиска аномальных значений в статистических данных, а также сравнение их с методом, предложенным автором.

Основные методы обнаружения выбросов

На сегодняшний день существуют следующие методы распознавания аномалий:

- статистический анализ;
- кластеризация;
- алгоритм ближайшего соседа;
- классификация;
- спектральные методы;
- гибридные методы.

При использовании статистического анализа определяется разница между построенной моделью и реальными данными. Если эта разница превышает определённый порог, то в данных существуют аномалии. Выделяют следующие группы методов статистического анализа:

- параметрические методы (на основе Гауссовой модели, на основе регрессионной модели, их комбинация);
- непараметрические методы (методы на основе гистограмм или функций ядра).

Кластеризация заключается в том, что все похожие экземпляры группируются в кластеры, если какой-либо экземпляр удален от центров кластеров более чем на определенную величину, то он считается аномальным. Также аномальными могут быть признаны разрозненные и незначительные кластеры.

В алгоритмах ближайшего соседа осуществляется определение расстояния или меры сходства между двумя экземплярами данных.

Метод классификации заключается в том, что наблюдения делятся на один или несколько классов, а те наблюдения, которые не принадлежат ни к одному из классов, признаются выбросами. Самыми распространенными подходами в этом методе являются:

- нейронные сети;
- Байесовы сети;
- метод на основе правил;
- метод опорных векторов.

При использовании спектрального метода на основе частотных характеристик данных строится модель, которая должна учесть большую часть изменчивости в данных [3-7].

В таблице 1 приведены примеры основных областей и решаемых задач, в которых применяется поиск выбросов, а также наиболее часто используемые методы применительно к каждой области.

Таблица 1 – Примеры применения методов поиска выбросов в различных областях

Область	Пример задачи	Метод
Медицина	вспышки заболеваний, отклонения в состоянии пациентов, ошибки записи	параметрические статистические методы, нейронные сети, байесовские нейронные сети, методы на основе правил, алгоритм ближайших соседей
Астрономия	отделение квазаров (активное ядро галактики) от звёзд	алгоритм ближайшего соседа
Компьютерные сети	обнаружение сетевых вторжений, взломов	все виды статистических методов, все виды классификации, кластеризация, ближайшего соседа
Обнаружение мошенничества	мошенничество с кредитными картами, мобильными телефонами, страховые агентства	статистические методы с использованием гистограмм, параметрические статистические методы, нейронные сети, методы на основе правил, кластеризация
Промышленность	поломки оборудования	параметрические и непараметрические статистические методы, нейронные сети, спектральный анализ
Торговля	выявление аномального спроса	параметрические статистические методы
Обработка изображений	спутниковые изображения, распознавание цифр, медицинские снимки	параметрические статистические методы, нейронные и байесовские сети, кластеризация, алгоритм ближайшего соседа

Как видно из таблицы, параметрические статистические методы используются для поиска аномальных значений в выборке практически во всех предметных областях.

Такое положение делает актуальной задачу совершенствования параметрических статистических методов поиска и обработки аномалий.

Сравнительный анализ статистических методов

Для сравнительного анализа эффективности статистических методов поиска аномалий в регрессии и предложенного автором метода были выбраны следующие методы:

- Эктона;
- Титьена-Мура-Бэкмана;
- Прескотта-Лунда;
- расстояние Кука;
- расстояние Махаланобиса.

Исходными данными для анализа является зависимость оборота розничной торговли непродовольственными товарами, млн. руб. от среднедушевого денежного дохода населения в Российской Федерации, руб./месяц. Для проверки работы методов, добавим в исходные данные четыре аномальных значения. В этом случае, значение коэффициента детерминации составляет 0,55.

При поиске аномальных значений методом Эктона было выявлено 3 подозрительных значения, как наибольшее отклонение исходных измерений от расчетных данных (e_i). После этого, по формуле (1) было рассчитано значение V, которое сравнивалось с критическим.

$$V = \frac{|e_k - \bar{e}|}{S_k}, \qquad (1)$$

где e_k – остаток предполагаемого выброса;

$\bar{e}$ – среднее по всем остаткам.

S_k – среднеквадратическое отклонение экспериментальных точек линии регрессии с учетом отбрасывания подозрительного наблюдения.

Остаток e_i с вероятностью α считается выбросом, если расчетное значение V больше критического V_α [8]. У двух выявленных подозрительных значений Y=46359 при X=29946 и Y=45644,07 при X=32285 расчетное значение V оказалось больше критического, поэтому эти значения признаются выбросами.

Далее для поиска аномального значения использовался метод Титьена-Мура-Бэкмана [9]. Он заключается в том, что по формуле (2) рассчитывается значение R_m. После этого, полученное значение R_m сравнивается с критическим значением Rα. Если полученное значение оказывается больше, то значение Y_i является выбросом.

$$Rm = \max\left|\frac{e_i}{S_i}\right|, \qquad (2)$$

где S_i – среднеквадратические отклонения остатков.

Используя формулу (2), было выявлено всего одно подозрительное значение. Величина, полученная по критерию, составила 2,71, однако она оказалась меньше критического значения равного 2,74 для уровня значимости 0,1, поэтому подозрительное значение выбросом согласно данного метода не является.

Третий метод – метод Прескотта-Лунда. С помощью данного метода по формуле (3) было получено значение R*=2,69 для подозрительного элемента, что является меньше критического равного 2,72. Следовательно, данное значение не признаётся выбросом.

$$R^* = \sqrt{n}\max\frac{|e_i|}{\sqrt{\sum_{i=1}^{n} e_i^2}}. \qquad (3)$$

Расстояние Кука представляет собой меру влияния определённых наблюдений на построенную регрессию. Для нахождения данной статистики чаще всего используется формула 4:

$$D_i = \frac{(\hat{y}_j - \hat{y}_{j(i)})}{pS_e^2 h_{ii}}, \qquad (4)$$

где $\hat{y}_j$ – ожидаемое значение регрессии (для j-го наблюдения), построенной по всей выборке;

$\hat{y}_{j(i)}$ – ожидаемое значение регрессии, построенной по выборке без i-го наблюдения;

p – число параметров модели (для линейной оно равно 2);

S_e^2 – среднеквадратическая ошибка модели, полученная при использовании всех данных;

h_{ii} – показатель влияния i-го наблюдения на коэффициенты модели. Представляет диагональные элементы матрицы проекции на пространство регрессоров $H=X(X^TX)^{-1}X^T$. Для парной линейной регрессии значение h_{ii} находится по формуле (5):

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \qquad (5)$$

Существуют различные подходы к определению выбросов с помощью расстояния Кука. Наиболее часто используется правило, что значение с расстоянием Кука D_i более 4/n (где n – количество наблюдений в выборке) считается выбросом.

Метод Кука является наиболее трудоёмким для ручного просчёта, поэтому поиск аномальных данных осуществлялся с использованием статистического пакета R. Было выявлено, что наибольшее влияние на модель оказывают наблюдения под номерами: 12, 19, 27. (рис. 1).

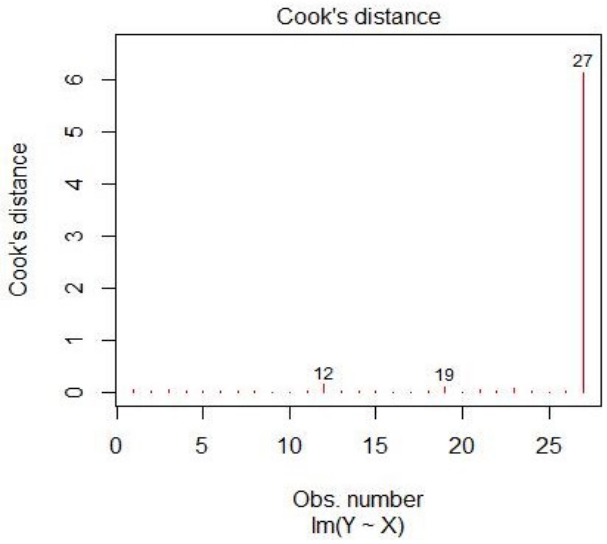


Рисунок 1 – Расстояние Кука

Эти наблюдения соответствуют следующим данным: Y=46262,16 при X=58848, Y=46359 при X=29946 и Y=45644,068 при X=32285.

Расстояние Махаланобиса определяет расстояние между двумя точками. Показывает, насколько значения наблюдений для независимых переменных отклоняются от среднего по всем наблюдениям. Расстояние Махаланобиса было рассчитано с помощью статистического пакета SPSS. В результате было найдено одно аномальное значение Y=46262,16 при X=58848.

Метод поиска аномалий, предложенный автором в [10, 11], основан на построении прямоугольной области надёжности, которая зависит от исходного уравнения регрессии. Те статистические данные, которые не попали в построенную область, признаются аномальными.

При применении данного метода были найдены все 4 аномальные измерения: Y=46262,16 при X=58848, Y=46359 при X=29946, Y=27597,16 при X=15800, Y=45644,068 при X=32285.

Результаты сравнения рассмотренных выше методов представлены в таблице 2.

Таблица 2 – Сравнительные данные методов поиска аномалий

Метод	Аномалии
Эктона	X=29946 Y=46359 X=32285 Y=45644,07
Титьена-Мура-Бэкмана	не выявлено
Прескотта-Лунда	не выявлено
Кука	X=58848 Y=46262,16 X=29946 Y=46359 X=32285 Y=45644,068
Махаланобиса	X=58848 Y=46262,16
Метод, предложенный в работе	X=58848 Y=46262,16, X=29946 Y=46359, X=15800 Y=27597,16, X=32285 Y=45644,068

Как видно из таблицы, методом поиска аномалий, предложенным автором было выявлено большее число аномалий, чем другими методами. Методом Эктона было обнаружено два аномальных наблюдения, а методами Титьена-Мура-Бэкманэ и Прескотта-

Лунда не было выявлено ни одного. Ближе всего по результативности к предложенному методу оказался метод Кука, однако данным методом были выявлены не все аномалии, а только 3, помимо этого количество элементарных операций возрастает при увеличении количества проверяемых подозрительных значений, а также отсутствует однозначный критерий того, какие из подозрительных значений признавать аномальными. Например, при использовании встроенной функции для расчёта расстояния Кука в статистическом пакете SPSS было выявлено 2, а не 3 аномальных значения. Таким образом, можно сделать вывод, что предложенный автором метод поиска аномалий является наиболее эффективным и быстродейственным.

Программная реализация предложенного метода

Для удобства использования, предложенного метода был разработан программный комплекс, которой состоит из взаимосвязанных приложений, написанных на языке C# и Visual Basic for Application для Microsoft Excel.

Пользователь может вводить данные двумя способами – вручную, непосредственно в самом приложении или загружать данные из Excel файла. Также, полученные результаты можно передать и сохранить в Excel. Вид стартового окна приложения представлен на рисунке 2. После ввода данных осуществляется их проверка на корректность и последующая сортировка.

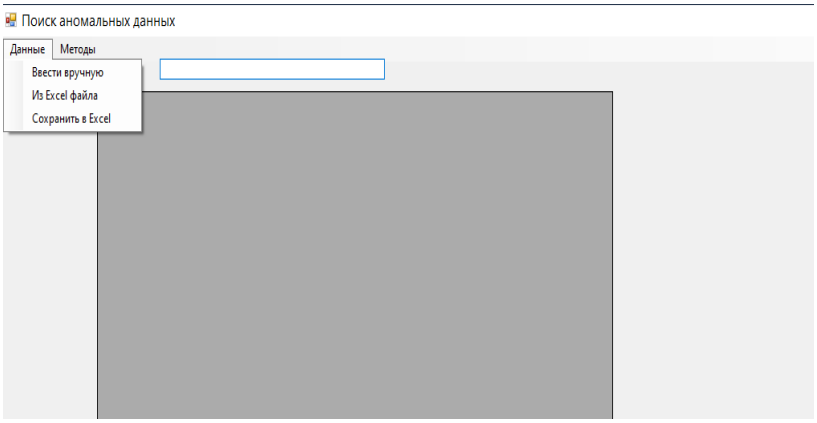


Рисунок 2 – Вид стартового окна программы

Далее пользователь может воспользоваться либо методом с последующим отбрасыванием аномальных данных, либо методом с корректировкой аномальных данных и их модификациями.

В зависимости от выбранного метода в программе рассчитываются показатели эффективности для каждого значения вероятности: коэффициенты детерминации R^2, величины доверительных интервалов, значения смещений, количество данных (исходное и после отбрасывания), точность, коэффициенты нового линейного регрессионного уравнения (рис. 3). Также, на соответствующей вкладке пользователь может посмотреть все найденные аномальные значения для различных вероятностей (рис. 4).

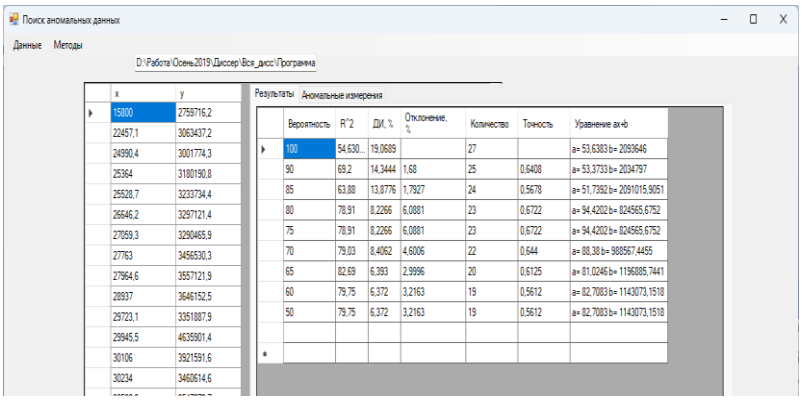
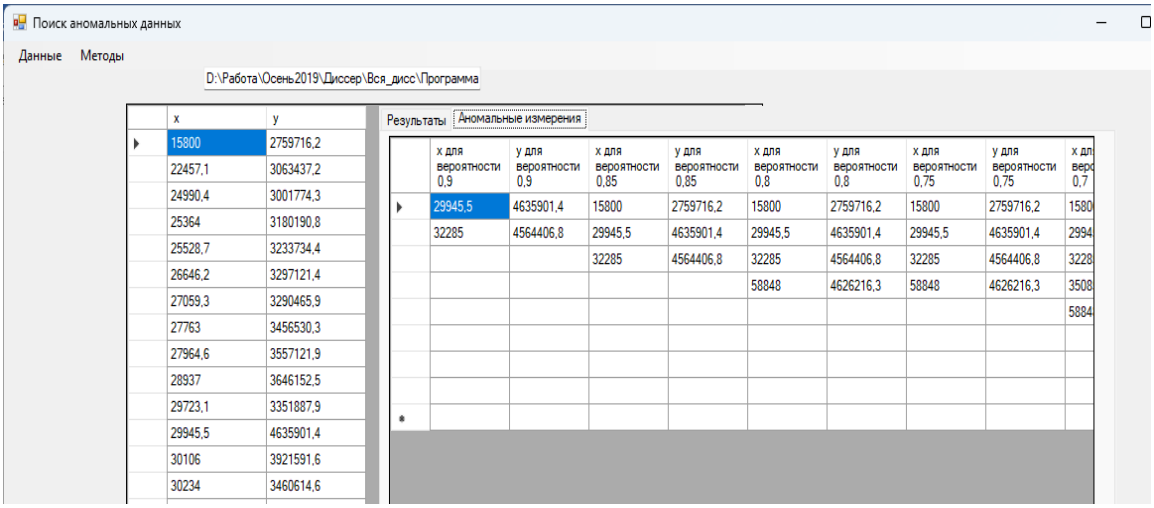


Рисунок 3 – Результаты работы программы



x	y
15800	2759716,2
22457,1	3063437,2
24990,4	3001774,3
25364	3180190,8
25528,7	3233734,4
26646,2	3297121,4
27059,3	3290465,9
27763	3456530,3
27964,6	3557121,9
28937	3646152,5
29723,1	3351887,9
29945,5	4635901,4
30106	3921591,6
30234	3460614,6

x для вероятности 0,9	y для вероятности 0,9	x для вероятности 0,85	y для вероятности 0,85	x для вероятности 0,8	y для вероятности 0,8	x для вероятности 0,75	y для вероятности 0,75	x дл вер 0,7
29945,5	4635901,4	15800	2759716,2	15800	2759716,2	15800	2759716,2	1580
32285	4564406,8	29945,5	4635901,4	29945,5	4635901,4	29945,5	4635901,4	2994
		32285	4564406,8	32285	4564406,8	32285	4564406,8	3228
				58848	4626216,3	58848	4626216,3	3508
								5884

Рисунок 4 – Аномальные данные

Разработанный программный комплекс позволяет быстро и удобно выявить аномальные данные в исходных статистических данных, устранить или откорректировать их, и получить результаты, на основе, которых можно осуществлять дальнейший анализ и прогнозирование.

Выводы

В статье рассмотрены статистические методы выявления аномальных данных, выполнен их сравнительный анализ. Помимо широкоизвестных методов в анализе использовался и разработанный автором метод. Анализ проводился на реальных данных. Наилучший результат показал метод, который был предложен в работе. С его использованием были выявлены все выбросы, содержащиеся в данных.

Для эффективного использования метода поиска аномалий, автором был разработан комплекс программ. Помимо поиска выбросов, в нём осуществляется последующая обработка выявленных аномальных измерений. Это позволяет получить адекватную модель, с использованием которой, можно строить более точные прогнозы.

Литература

1. Караулова, А.В. Применение регрессионного анализа при решении реальных задач технического характера / А. В. Караулова, И. П. Базилевский // «Молодая наука Сибири»: электрон. науч. журн. – 2020. – №3(9). - Режим доступа: http://mnv.irgups.ru/toma/39-2020.

2. Копырин, А.С. Оценка влияния аномалий на результаты анализа массивов экономических данных / А. С. Копырин, Е. В. Видищева // Modern Economy Success, 2021. - № 2. - С. 235–240.

3. Chandola, V., Banerjee, A., and Kumar, V. Anomaly detection: A survey. ACM Comput. Surv. – 41, 3, Article 15 (July 2009) – 58 pages.

4. Wilson J. Holton, Keating Barry P., Beal Mary Regression Analysis: Understanding and Building Business and Economic Models Using Excel, 2nd Edition. — New York, USA, Business Expert Press, LLC, 2016. — 205 p.

5. Кириченко, А. В. (и др.) Математические модели и методы анализа и прогнозирования: предварительная обработка результатов эксперимента, проверка статистических гипотез, корреляционный анализ, парный регрессионный анализ: учебное пособие. - Саратов: КУБиК, 2019. - 259 с.

6. Кузовлев, В.И., Орлов А.О. Метод выявления аномалий в исходных данных при построении прогнозной модели решающего дерева в системах поддержки принятия решений // Наука и образование. МГТУ им. Н.Э. Баумана. Электрон. журн. 2012. № 9. URL: https://cyberleninka.ru/article/n/metod-vyyavleniya anomaliy-v-ishodnyh-dannyh-pri-postroenii-prognoznoy-modeli-reshayuschego-dereva-v-sistemah-podderzhki-prinyatiya/viewer.

7. Девянин, И.С. Предварительная обработка данных для машинного обучения // Фундаментальные и прикладные исследования в физике, химии, математике и информатике, 2021. - С. 117–121.

8. Попукайло, В.С. Обнаружение аномальных измерений при обработке данных малого объема // Технология и конструирование в электронной аппаратуре, 2016. – № 4-5 – С. 42-46.

9. Кобзарь, А. И. Прикладная математическая статистика. Для инженеров и научных работников [Текст] / А. И. Кобзарь. – М.: ФИЗМАТЛИТ, 2012. – 816 с.

10. Рычка, О.В. Разработка алгоритма реализации методов повышения качества регрессионных моделей, используемых при проектировании технических систем. // Научный журнал «Информатика и кибернетика». – Донецк: ДонНТУ, 2020. - № 3 (21). - С.42-48.

11. Рычка, О.В. Анализ эффективности усовершенствованных методов поиска и обработки аномалий для нелинейных моделей с внутренней линейностью // Международный рецензируемый научно-теоретический журнал «Проблемы искусственного интеллекта». – Донецк, 2020. – Вып. №3(18) – С. 101-110.

Рычка О.В. Сравнительный анализ методов интеллектуальной обработки данных для повышения качества прогнозных моделей*. В данной статье отмечена важность предварительной обработки в анализе данных. Описаны результаты сравнительного анализа эффективности различных методов поиска аномальных значений в статистических данных и предложенного автором метода. Представлена программная реализация предложенного метода. Анализ выполнялся на реальных данных. Реализованный в работе метод обнаружения и обработки выбросов позволяет определять более точные значения различных показателей и способствует построению достоверных прогнозов.*

Ключевые слова: *аномальные измерения, выброс, поиск аномалий, программный комплекс, сравнительный анализ, прогноз.*

Rychka O.V. Comparative analysis of intelligent data processing methods to improve the quality of predictive models. *This article highlights the importance of preprocessing in data analysis. It describes the results of a comparative analysis of the effectiveness of various methods of searching for anomalous values in statistical data and the method proposed by the author. A software implementation of the proposed method is presented. The analysis was performed on real data. The method of detection and processing of outliers implemented in the work makes it possible to determine more accurate values of various indicators and contributes to the construction of reliable forecasts.*

Key words: *anomalous measurements, outlier, search for anomalies, software package, comparative analysis, forecast.*

Статья поступила в редакцию 24.05.2023
Рекомендована к публикации профессором Зори С. А.