УДК 004.89

Е. В. Видищева
ФБГОУ ВО «Сочинский государственный университет», Сочи, e-mail: evgenia-vv@mail.ru

А. С. Копырин
ФБГОУ ВО «Сочинский государственный университет», Сочи, e-mail: kopyrin_a@mail.ru

М. С. Василенко
ФБГОУ ВО «Сочинский государственный университет», Сочи, e-mail: renxi@yandex.ru

АНАЛИЗ И УТОЧНЕНИЕ КЛАССИФИКАЦИИ АНОМАЛИЙ И ВЫБРОСОВ НА ЭКОНОМИЧЕСКИХ ДАННЫХ

Ключевые слова: интеллектуальный анализ данных, аномалии, выбросы, экономические данные.

В эпоху информатизации интеллектуальный анализ данных применяется практически во всех сферах человеческой деятельности. Достоверность результатов интеллектуального анализа данных напрямую определяет качество внутренних и внешних процессов, снижает вероятность непредвиденных ситуаций, позволяет составлять точные модели процессов и делать реалистичные прогнозы. В связи с этим особую актуальность обретает обнаружение и нейтрализация отклонений в исходных данных. Данная работа посвящена изучению существующих классификаций аномалий и выбросов на данных экономического характера. На сегодняшний день научная база по исследованию интеллектуального анализа экономических данных крайне ограничена. В работе рассмотрены различные классификационные подходы к аномальным элементам в данных, приведены примеры выбросов на экономических данных и определена важность своевременного обнаружения и устранения выбросов для получения достоверного результата. Целью исследования является анализ существующей теоретической базы по интеллектуальному анализу данных и оценка возможности ее применения к данным экономического характера. В результате исследования были выявлены классификационные признаки и произведена группировка существующих классификаций. Анализ работ по исследуемой тематике также позволит дополнить научную базу новой классификационной группой.

E. V. Vidishcheva
Sochi State University, Sochi, e-mail: evgenia-vv@mail.ru

A. S. Kopyrin
Sochi State University, Sochi, e-mail: kopyrin_a@mail.ru

M. S. Vasilenko
Sochi State University, Sochi, e-mail: renxi@yandex.ru

ANALYSIS AND CLARIFICATION OF ANOMALIES AND OUTLIERS CLASSIFICATIONS IN ECONOMIC DATA

Keywords: data mining, anomalies, outliers, economic data.

In the time of informatization data mining is used in almost all spheres of human activity. The reliability of data mining results directly determines the quality of internal and external processes, reduces the possibility of unforeseen situations, and allows creating accurate models of processes and making realistic predictions. In this regard, the detection and neutralization of deviations in the original data becomes particularly relevant. This work is devoted to the study of existing classifications of anomalies and outliers in economic data. To date, the scientific base for the study of intellectual analysis of economic data is extremely limited. The paper considers various classification approaches to the anomalous elements in the data, provides examples of emissions in economic data and determines the importance of timely detection and elimination of emissions to obtain reliable results. The aim of the study is to analyze the existing theoretical base for data mining and assess the possibility of its application to economic data. As a result of the study, classification features were identified and the existing classifications were grouped. The analysis of works on the subject under study also allows complementing the scientific base with a new classification group.

Введение

Аномалии при анализе данных создают помехи и сказываются на достоверности информации. Аналитика экономических данных подразумевает обработку крупного массива данных, полученных путем измерений, опросов или экспертных оценок. Интеллектуальный анализ

экономических данных позволяет описывать процессы и явления, создавать модели и прогнозы будущего развития. Экономические модели используются как на микро, так и на макроуровне, позволяют прогнозировать вероятность банкротства, финансовые временные ряды и прочие экономические индикаторы. Результат измерения, существенно выбивающийся из подборки, может серьезно исказить итоговую оценку. Именно поэтому крайне важно различать возможные типы и формы возникновения аномальных элементов для их своевременного обнаружения и нейтрализации.

Цель исследования

Целью исследования является обзор отечественных и зарубежных подходов к классификации аномальных явлений и выбросов на данных, а также оценка применимости существующих классификаций к данным экономического характера.

Материал и методы исследования

В ходе исследования использовались материалы из зарубежных и российских периодических изданий, и монографий, а также общедоступные ресурсы сети Интернет. Для достижения поставленных целей были применены эмпирические и теоретические методы исследования, такие как сбор, изучение и анализ данных, обобщение, сравнение и классифицирование.

Результаты исследования и их обсуждение

Исследованию аномалий и выбросов, возникающих в процессе интеллектуального анализа данных, посвящены работы как российских, так и зарубежных ученых. Причем понимание и противопоставление исследуемых понятий различно. Одни авторы употребляют понятия аномалии и выброса в качестве синонимов, другие разделяют дефиниции. В зарубежной специализированной литературе преимущественно применяется понятие выброс, тогда как подавляющее большинство отечественных исследований посвящены изучению аномалий. В рамках данного исследования данные понятия носят синонимичный характер.

Стоит отметить, что научная база по изучению аномалий и выбросов в экономических данных крайне ограничена. Среди всего многообразия тематических работ лишь единицы посвящены исследованию аномалий в данных экономического характера (Толви Д., 2001; Минтс А., 2017) [7, 10].

На сегодняшний день не существует общепринятой классификации аномальных явлений или выбросов на данных. Наиболее часто в работах отечественных и зарубежных ученых встречается классификация, в рамках которой выделены три типа аномалий: точечные, контекстные и коллективные (Каранжит Сингх, Шучита Упадьяя, 2012; Чандола В., Банерджи А., Кумар В., 2009) [4, 9]. Примеры данной типологии применительно к экономическим данным представлены на рис. 1–3.

В упомянутой классификации аномалии разделены по форме возникновения. Точечная аномалия представляет собой отдельный экземпляр данных, который не вписывается в общую картину и является аномальным по отношению к остальным данным. Точка А (рис. 1), размещенная на совокупности данных о соотношении объема выпускаемой продукции и объема капиталовложений отдельного предприятия, является примером точечной аномалии на экономических данных.

Второй тип – контекстные аномалии также называют условными, так как признак аномальности проявляется только в рамках определенного контекста. В отличие от точечной аномалии, выявление контекстной аномалии обусловлено наличием поведенческих и контекстных атрибутов. В качестве примера данного типа аномалии (точка В) представлены данные о спросе на туристические услуги – численности иностранных туристов на территории города Сочи (рис. 2). Для города Сочи точка В не является аномальной только благодаря наличию контекстных атрибутов, а именно проведение в 2014 году международного мероприятия в регионе. Для любого другого города России подобное значение было бы расценено как аномальное.

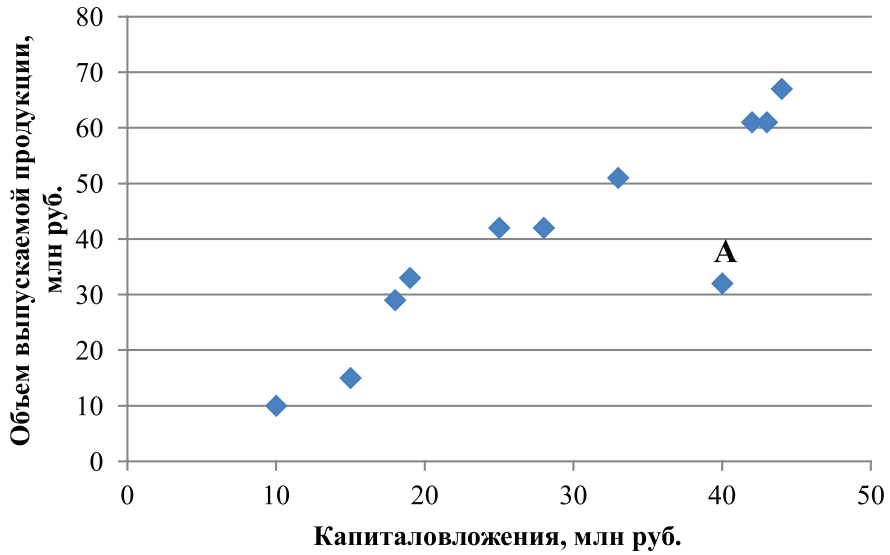

Рис. 1. Пример точечной аномалии на экономических данных

Рис. 2. Пример контекстной аномалии на экономических данных

К следующему типу аномалий – коллективным относят совокупности взаимосвязанных данных, являющихся аномальными по отношению ко всему набору данных. Причем, в отдельности каждый элемент коллективной аномалии не является аномальным, лишь их совместное появление рассматривается аномально. Точкой С отмечена коллективная аномалия в наборе данных об объеме продаж определенной марки автомобиля в зависимости от динамики цены (рис. 3). Единичный рост спроса при росте цены не является аномальным, так как может быть обусловлен экономической ситуацией или прочими внешними условиями. А повторение аналогичной ситуации на протяжении трех отчетных периодов является аномальным.

Группа американских ученых под руководством Камбера М. предлагает несколько другой подход к данной классификации, определяя простейшие выбросы (элементы данных, значительно отличающиеся от остальной части набора данных) как глобальные [6].

Кришна Моди дополняет традиционную классификацию, разделяя выбросы на реальные и ошибочные (призрачные) [8]. Реальными автор называет выбросы, которые действительно содержат в себе нетипичную, а возможно и ценную

информацию – нечто новое и инновационное. Их устранение полностью стабилизирует информацию, но при этом может стать препятствием при обнаружении уникальной тенденции. Призрачные выбросы при интеллектуальном анализе данных возникают в связи с внутренними проблема или сбоями и заключаются в ошибочном определении той или иной совокупности данных как аномальных.

Наиболее обширная из существующих классификаций представлена в работе Ральфа Фуртуса [5]. Классификация основана на пересечении двух классификационных признаков: тип информации и мощность связи (рис. 4).

Аномалии 1 типа – экстремальных значений подразумевают возникновение чрезвычайно высокого или низкого показателя в совокупности данных. При анализе экономических данных этот вид аномалии встречается достаточно часто, и для его обработки используется показатель стандартного отклонения. Многомерные аномалии зависят от нескольких атрибутов, и для их выявления необходимо проводить совместный анализ как минимум по двум признакам (атрибутам). Всего в классификации представлено 6 видом аномалий.

Одномерный тип связи означает, что аномалия возникает в рамках одного измерения, многомерные выбросы выделяются сразу в нескольких измерениях. Используя исключительно критерий мерности или охвата, можно также разделить выбросы на несколько групп: одномерные, многомерные и категориальные (возникают в отдельных категориях данных) [3].

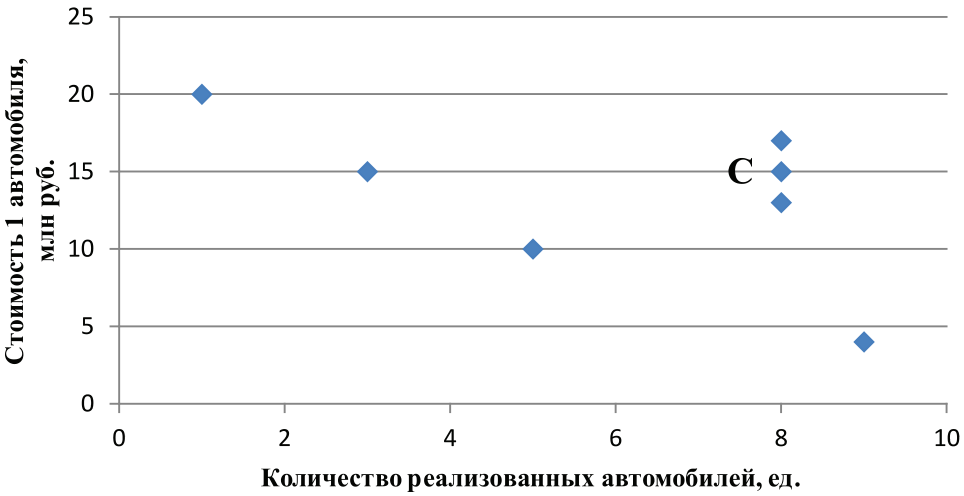

Рис. 3. Пример коллективной аномалии на экономических данных

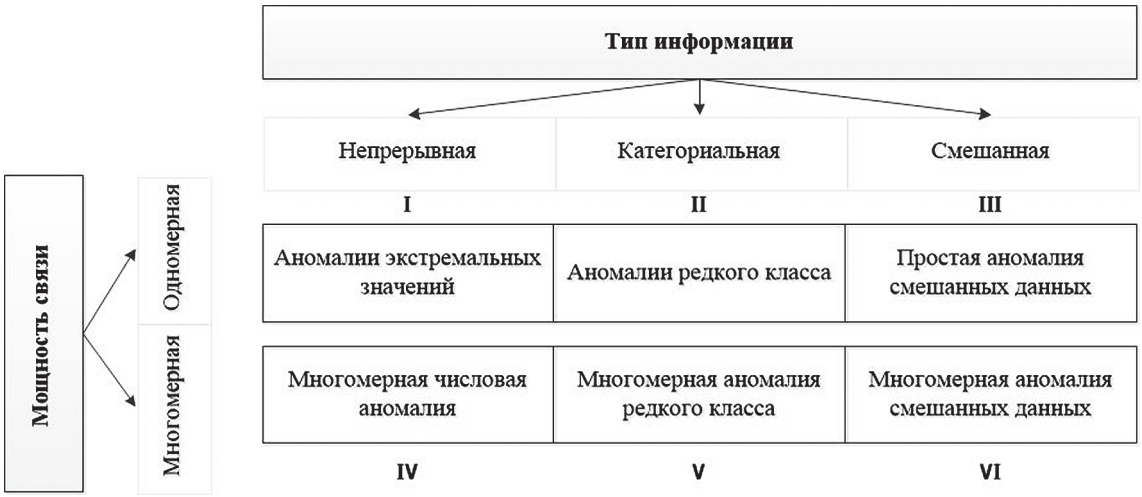

Рис. 4. Классификация аномалий Р. Фуртуса

Среди исследований, посвященных изучению аномальных явлений и выбросов непосредственно на экономических данных, следует выделить работу Дж. Толви [10]. Автор рассматривает три типа выбросов на макроэкономических данных, среди них:

1. Аддитивные выбросы – слишком большое или маленькое значение, единожды встреченное в выборке.

2. Выброс временного изменения – выброс, воздействие которого на общую совокупность данных постепенно угасает, и ряд возвращается к обычному уровню.

3. Выброс сдвига уровня – выброс, оказывающий перманентное воздействие на все последующие элементы выборки, то есть приводит к изменению уровня данных.

Рассмотренные выше классификации можно назвать унифицированными и применить как к данным экономического характера, так и к любому другому набору данных. Также выбросы могут быть разделены по силе и продолжительности воздействия, по источнику возникновения и охвату. В таблице представлены существующие классификации выбросов и аномалий на данных с выделением классификационного признака.

Источниками формирования данных экономического характера служат административные и статистические ресурсы, данные внутреннего учета предприятий, экспертные оценки и опросы. Вероятность возникновения ошибки, образующей в дальнейшем аномальное значение в наборе данных, при сборе исходной информации крайне велика. Причем выброс может быть сформирован как искусственным, так и случайным путем. Искусственные выбросы появляются в связи с неверным предоставлением информации, типографическими ошибками, умышленной недостоверностью данных, либо ошибочно сформированной выборкой. Случайные выбросы связаны с выбором конкретного образца данных из выборки. Присутствие любого из данных выбросов может серьезно повлиять на результаты аналитического исследования. Однако стоит отметить, что практические исследования, представленные в литературе, подтверждают существование выбросов,

Обзор существующий классификаций аномалий и выбросов

Классификационный признак	Авторы	Типы аномалий/выбросов
Форма возникновения	Чандола В., Банерджи А., Кумар В., Сайн К., Упадьяя Ш., Хан Дж., Камбер М., Пей Дж., Моди К., Оза Б.	Точечные, контекстные (условные), коллективные (глобальные)
Сущность выброса	Моди К., Оза Б.	Реальные, ошибочные (призрачные)
Тип информации и мощность связей	Фуртиус Р.	Аномалии экстремальных значений, редкого класса, простая аномалия смешанных данных, многомерная числовая аномалия, многомерная редкого класса, многомерная смешанных данных
Сила воздействия	Аггарвал С.С.	Слабый выброс, сильный выброс
Продолжительность воздействия	Хуберт М., Руссо П., Сигарт П.	Изолированные выбросы, постоянные выбросы (сдвиговые, амплитудные и выбросы формы)
Стадия возникновения	Браун Г.	Выбросы в данных опроса, в административных данных, в моделировании
Источник	Анскомб Ф.Дж.	Искусственные, случайные
Охват	Богарт З., Роббинс Дж.	Одномерные, многомерные, категориальные
Форма влияния	Толви Дж.	Аддитивные, временного изменения, сдвига уровня

Источник: составлено автором на основе [1–10].

не сказывающихся на общей совокупности данных. К примеру, группа ученых во главе с Алварез Е. при анализе показателей бедности ряда стран пришла к выводу, что наличие выбросов не сказалось на результатах оценки, и после их удаления существенного изменения в данных не произошло. Таким образом, можно разделить выбросы по наличию воздействия на набор данных: искажающие и нейтральные.

Выводы (заключение)

Изучение существующих подходов к классификации аномальных явлений и выбросов позволяет заключить, что большинство классификационных групп носят унифицированный характер. Следовательно, разнообразные типы выбросов могут быть обнаружены в данных любого характера, в том числе экономического. Львиную долю существующей научно-исследовательской базы по изучаемой тематике составляют труды зарубежных ученых. Российский вклад в изучение интеллектуального анализа данных не столь обширен. Также отличительной чертой зарубежных подходов является использование преимущественно понятия выброс, тогда как в российской практике чаще встречается термин аномалия.

В ходе исследования существующие классификации были изучены и сгруппированы по выявленным классификационным признакам. Также, основываясь на проблематике изученных тематических исследований, был выявлен еще один признак дифференциации – наличие воздействия на выборку. И исходя из этого предложена классификация выбросов по двум типам: искажающие совокупность данных и нейтральные по отношению к ней.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-01-00370.

Библиографический список

1. Шкодырев В.П., Ягафаров К.И., Баштовенко В.А., Ильина Е.Э., Обзор методов обнаружения аномалий в потоках данных, 2017 [Электронный ресурс]. URL: https://docplayer.ru/58831564-Obzor-metodov-obnaruzheniya-anomaliy-v-potokah-dannyh.html (дата обращения: 13.05.2019).

2. Anscombe F.J., Rejection of outliers. Technometrics, 1960. Vol. 2, №2, P. 123–147.

3. Bogart Zach, Robbins Joyce, Everything you need for Exploratory Data Analysis & Visualization, 2019-04-17. [Электронный ресурс]. URL: https://edav.info/outliers.html (дата обращения: 13.05.2019).

4. Chandola V., Banerjee A., Kumar V., Anomaly detection: A survey. ACM Computing Surveys, 2009. Vol. 41 (3), 58 p.

5. Foorthuis Ralph, A Typology of Data Anomalies. 7 International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Session on Fuzzy methods in Data Mining and Knowledge Discovery. Springer 2018, 13 p. [Электронный ресурс]. URL: https://tunguska.home.xs4all.nl/Publications/Docs/A%20Typology%20of%20Data%20Anomalies%20-%20Foorthuis%20-%20IPMU%202018.pdf (дата обращения: 16.05.2019).

6. Han Jiawei, Kamber Micheline, Pei Jian, Data Mining Concepts and Techniques: Third Edition. Published by Morgan Kaufmann, 2011 [Электронный ресурс]. URL: https://learning.oreilly.com/library/view/data-mining-concepts/9780123814791/ (дата обращения: 16.05.2019).

7. Mints Alexey. Classification of tasks of data mining and data processing in the economy. Baltic Journal of Economic Studies, 2017. Vol. 3, №3. P. 47–52.

8. Modi Krishna, Prof Oza Bhavesh, Outlier Analysis Approaches in Data Mining. International Journal Of Innovative Research In Technology, 2016. Vol. 3, Issue 7. P. 6–12.

9. Singh Karanjit, Upadhyaya Shuchita, Outlier Detection: Applications And Techniques. International Journal of Computer Science Issues, 2012. Vol. 9, Issue 1, №3. P. 307–323.

10. Tolvi Jussi, Outliers in Eleven Finnish Macroeconomictime Series. Finnish Economic Papers, 2001. Vol. 14, №1. P. 14–32.