DOI: 10.34185/1991-7848.itmm.2021.01.029

ІДЕНТИФІКАЦІЯ ТА ОБРОБКА АНОМАЛІЙ ДАНИХ В ЗАДАЧАХ МАШИННОГО НАВЧАННЯ

Калініна І.О., к.т.н., доцент,

Гожий О.П., д.т.н., професор,

Чорноморський національний університет ім.П.Могили, Україна.

Вступ. На етапі підготовки даних до моделювання в процедурах машинного навчання, при формуванні початкової вибірки виникають проблеми наявності частини даних, які відрізняються від загальної вибірки та знаходяться на статистично далекій відстані. Такі дані називаються аномаліями або викидами. Аномалії або викиди – це дані, які суттєво відрізняються від інших спостережень [1]. Вони можуть відповідати реальним відхиленням, але можуть бути і просто помилками. Викиди з'являються у вибірках даних з різних причин. Вони можуть бути наслідками: помилок в даних (неточності вимірювання, округлення, невірного запису і т.п.); наявності шумових об'єктів (невірно класифікованих об'єктів); наявності об'єктів «інших» вибірок (наприклад, показання датчика, який вийшов з ладу). У якості справжніх викидів визначимо викиди «в широкому сенсі», тобто дані з набору, які спотворюють границі класу / кластеру. Викиди бувають не тільки в табличних (структурованих) даних, вони можуть бути в графах, часових рядах і т.д.

Основний матеріал. Для ідентифікації та обробки аномалій пропонується процедура схема, якої представлена на рис.1. Процедура складається з трьох етапів. На першому етапі виявляються викиди у вибірках даних. Серед особливостей цього етапу слід відмітити велику кількість методів, які можна застосувати на цьому етапі. Вибір конкретного методу відбувається в залежності від задачі машинного навчання, структури набору даних і типів даних, які обробляються. Серед методів, які використовуються на цьому етапі слід відмітити наступні: методи статистичних тестів (метод Z-оцінки, метод Kurtosis measure, ESD-тест та інш.), методи метричніх тестів (метод KNN, метод LOF), методи модельних тестів, ітераційні методи, методи машинного навчання (SVM, випадкового лісу), ансамблеві методи [2,3].

Головна особливість другого етапу – це те, що на цьому етапі проводиться аналіз причин появи викидів. До причин появи викидів можна віднести: причини пов'язані з похибками у вимірюванні та причини пов'язані з похибками обробки даних, результати зовнішніх впливів, або помилки в записах даних. В залежності від результатів аналізу на цьому етапі можливе повернення до першого етапу з метою вибору іншого методу або для корегування критерію відбору даних.

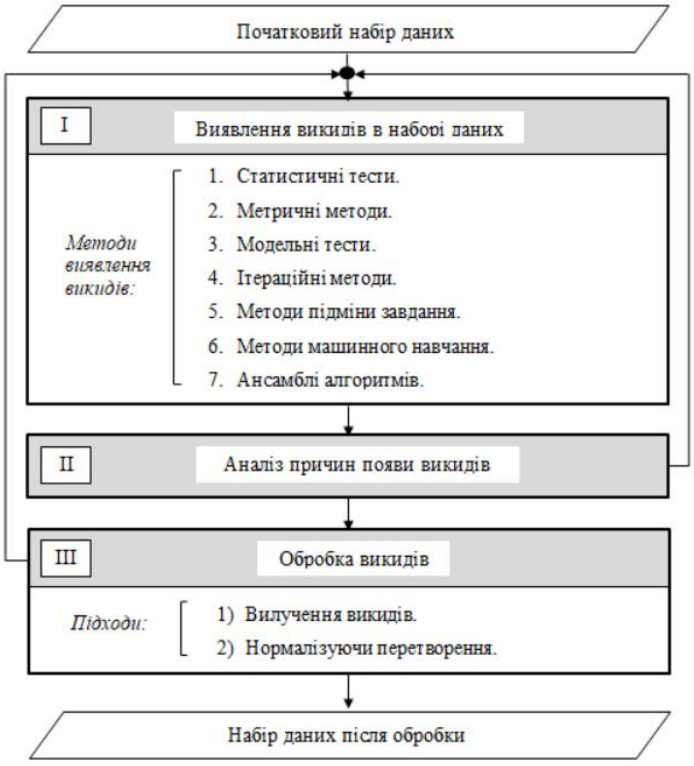

Рисунок 1 – Схема процедури ідентифікації та обробки даних

На третьому етапі здійснюється остаточна обробка наборів даних з викидами, в яких відбувається видалення викидів або нормалізуючі перетворення. Якщо результати не влаштовують, тоді необхідно повернутись до етапу ідентифікації, обрати інший метод та повторно перевірити набір даних на наявність викидів.

При виявлені даних з початкового набору, що є підозрілі на аномалії, їх можна вилучити з набору. Але таке рішення – є суб'єктивним рішенням. Важливо пам'ятати про те, що причини для виникнення подібних спостережень можуть бути різними. Так, видалення викидів, що виникли через невдале планування експерименту або не якісного вимірювання, може бути

іноді цілком виправданим. У той же час, незвичайні спостереження серед значень залежної змінної можуть вимагати більш точного підходу, особливо якщо вони відображають природну варіабельність цієї змінної.

Альтернативою видалення незвичайних значень вибірки є нормалізуючі перетворення, найчастіше, логарифмування. У загальному випадку, знайти оптимальне рішення дозволяє перетворення Бокса-Кокса. Універсальне сімейство перетворень Бокса-Кокса випадкової величини x є ступеневим перетворенням $x' = \frac{x^\lambda - 1}{\lambda}$ з довільним додатним або від'ємним показником ступеня λ. Оскільки ділення на нуль призводить до невизначеності, то при $\lambda = 0$ використовується логарифмічна перетворення $x'(\lambda) = \ln(x)$. Значення λ можна знайти, наприклад, за допомогою максимуму логарифма функції максимальної правдоподібності. Використання подібних перетворень значно зменшує кількість підозрілих на викиди значень в наборі даних.

Висновки. Таким чином, вихідні значення залежної змінної зазвичай представляють інтерес при побудові регресійних моделей в задачах прогнозування та їх перетворення може порушити змістовний сенс перевірки гіпотез. Наприклад, виконавши перетворення, яке нормує дисперсію, виявляється, що щомісячний дохід олігарха мало відрізняється від доходу пенсіонера. Тому краще підібрати метод аналізу, який заснований на розподілі ймовірностей, якій допускає асиметрію розкиду значень в області екстремального значення великих даних, наприклад, гамма-розподіл для безперервних змінних або розподіл Пуассона для дискретних кількісних змінних. Остаточне рішення про видалення аномальних значень з набору даних приймається в залежності від особливостей конкретної задачі.

Література

1. Chandola V., Banerjee A., Kumar V. Anomaly detection: A survey //ACM computing surveys (CSUR). – 2009. – Т. 41

2. Chandola V., Banerjee A., Kumar V. Anomaly detection for discrete sequences: A survey //IEEE Transactions on Knowledge and Data Engineering. – 2012. – Т. 24. – №5. – С. 823-839

3. Aggarwal C. C., Sathe S. Outlier Ensembles: An Introduction. – Springer, 2017.

IDENTIFICATION AND PROCESSING OF DATA ANOMALIES IN MACHINE LEARNING TASKS

Kalinina Iryna, Gozhyj Oleksandr

Abstract. The paper presents the procedure of identification and processing of data anomalies at the stage of preliminary data processing in machine learning tasks. The procedure consists of three stages. At the first stage, emissions are detected in the data samples. A large number of methods are used for this. The choice of a particular method depends on the task of machine learning, the structure of the data set and the types of data being processed. The methods used at this stage are methods of statistical tests, methods of metric tests, methods of model tests, iterative methods, methods of machine learning, ensemble methods. Until the second stage, the analysis of the causes of emissions is carried out. The causes of emissions include: causes of measurement errors and causes of data processing errors, the results of external influences, or errors in data records. In the third stage, there is a final processing of data sets with emissions, in which there is a removal of emissions or normalizing transformations. The effectiveness of the procedure was tested on different data sets.

Keyword. Data anomalies, preliminary data processing, tests, machine learning tasks, data sets.

References

1. Chandola V., Banerjee A., Kumar V. Anomaly detection: A survey //ACM computing surveys (CSUR). – 2009. – T. 41

2. Chandola V., Banerjee A., Kumar V. Anomaly detection for discrete sequences: A survey //IEEE Transactions on Knowledge and Data Engineering. – 2012. – T. 24. – №5. – C. 823-839

3. Aggarwal C. C., Sathe S. Outlier Ensembles: An Introduction. – Springer, 2017.