УДК 004.056
В.В. Баранніков
Тернопільський національний технічний університет імені Івана Пулюя, Україна

ОСОБЛИВОСТІ ЗАВДАННЯ ВИЯВЛЕННЯ АНОМАЛІЙ

V.V. Barannikov

FEATURES OF ANOMALIES DETECTION TASK

Аномалії — це закономірності даних, котрі не відповідають добре визначеному поняттю нормального поведінки. Проблема детектування цих патернів називається виявленням аномалій (ВА). Важливість ВА зумовлена тим фактом, що аномалії даних приводять до значної і дієвої інформації в різних областях застосування. Наприклад, ненормальні схеми трафіку в комп'ютерних мережах можуть означати, що комп'ютери відправляють конфіденційні дані в несанкціоновані місця призначення, відхилення в даних транзакції по кредитній картці можуть вказувати на кредитні картки або посвідчення особи і т.п.

Навіщо взагалі шукати аномалії?

По-перше, щоб покращити якість моделі. Тобто це завдання передобробки даних.

По-друге, в даних можуть бути шуми. Таким чином, аномалії зашумлюють наші дані і через ці викиди наш алгоритм може перенавчитися та видавати невірні оцінки. Тобто, є мета уникнення подальшого перенавчання.

По-третє, вивчення викидів. Можливо у деякій системі з платною підпискою є кілька аномальних користувачів, які платять у десятки разів більше за всіх інших. Тоді нам потрібно вивчити, що це за користувачі і що потрібно зробити, щоб їх зберегти. На основі цього вже вирішувати, чи варто ці аномалії виключати із даних, чи ні.

По-четверте, виявлення поломок. Зазвичай цим займаються диспетчери, які сотні годин бачать графіки зміни показів тих чи інших приладів і в деяких випадках їм вдається запобігти поломці. Або після її виникнення виявити, де саме і коли вона сталася.

Якщо доручити це й аналогічні завдання алгоритму, то, можливо, поломок більше не виникатиме, оскільки відсутній людський фактор неуважності. Система точно і беззастережно визначить, чи ця подія є аномалією, чи ні. Однак навіть алгоритм може помилятися, але це вже питання правильного вибору та складання моделі.

У більшості випадків аномальні дані, що визначаються нами, відносяться до виявлення викидів. Після навчання цих даних ми шукатимемо аномальні точки в новому наборі даних.

Виявлення новизни (ВН) - це метод, що дозволяє ідентифікувати нові чи невідомі шаблони та закони даних. Передумова виявлення новизни у тому, що набір навчальних даних, як відомо, є «чистим» і не забруднений реальними «шумовими» даними чи реальними «викидами», та був після навчання цих даних нові дані навчаються для пошуку шаблонів даних новизни. ВН, в основному, застосовується для дослідження та розпізнавання нових шаблонів, тем і тенденцій, включаючи обробку сигналів, комп'ютерний зір, розпізнавання образів, інтелектуальних роботів та інші технічні вказівки, а також сфери застосування, такі як дослідження потенційних захворювань, відкриття нових видів, надбання нових тем спілкування тощо. ВН пов'язані з ВА.

На початку точки новизни часто з'являються у даних стороннім чином. Цей сторонній спосіб зазвичай сприймається як сторонній. Тому шаблони виявлення та розпізнавання цих двох типів дуже схожі. Однак через деякий період часу, коли дані новизни підтверджуються як нормальний патерн, наприклад, нове захворювання ідентифікується як поширене захворювання, патерн новизни буде об'єднаний у нормальний патерн і більше не буде ставитись до категорії аномальних точок.