УДК 004.67

ДОСЛІДЖЕННЯ МЕТОДІВ АНАЛІЗУ АНОМАЛІЙ В DATA MINING

Сергієнко Д.В.
Науковий керівник – ст. викл. Климова І.М.
Харківський національний університет радіоелектроніки, каф. СТ
м. Харків, Україна
тел.: +38(095) 804-48-91, email: dmytro.serhiienko1@nure.ua

This work is devoted to the study of anomaly analysis. Anomaly analysis is an important part of data mining, as it allows to identify unusual or outlier data that may be important for decision making in a particular situation. In addition, anomaly analysis can be used as a standalone data analysis method and can also be incorporated into more complex analytical processes such as clustering or classification. Anomaly analysis can help identify important information that can be used for decision-making in various fields of activity, such as business, medicine, science, etc.

В загальному випадку, аномалія визначається як відхилення від очікуваної нормальної поведінки. Для виявлення аномалій, можна визначити область, яка представляє нормальну поведінку і оголосити будь-які спостереження, що не належать до цієї області, як аномалії.

Серйозною проблемою будь-якого автоматичного виявлення явно неправильних даних полягає в тому, що правильна модель може бути викинута разом з некоректними даними. Без консультації з людиною-експертом, неможливо визначити, чи є конкретний випадок помилкою, чи чи він просто не відповідає типу моделі, яка застосовується. Наприклад, у статистичній регресії допомагає візуалізація. Зазвичай навіть неспеціалісту буде візуально очевидно, якщо дані виділяються з загальної тенденції. Проте більшість проблем класифікації не можна так легко візуалізувати. Причиною цього є поняття «типу моделі», яке є більш тонким, ніж лінія регресії. Навіть якщо на більшості стандартних наборів даних можна отримати хороші результати, відкинувши випадки, які не підходять під модель дерева рішень, це не обов'язково буде дуже зручним, коли йде мова про справу з конкретним новим набором даних. Існує ризик, що, новий набір даних просто не підходить для змодельованого дерева рішень. Щоб зменшити цю ймовірність, можна скористатися наступним підходом: Використовувати декілька різних схем навчання (наприклад, дерево рішень, метод найближчого сусіда та лінійна дискримінантна функція та ін.) для фільтрації даних. Недолік прямого використання декількох схем полягає в тому, що всі схеми не можуть класифікувати екземпляр правильно, перш ніж він буде визнаний помилковим і видалений з даних. У деяких випадках фільтрація даних у такий спосіб і використання відфільтрованих даних як вхідних даних для остаточної схеми навчання дає кращі результати, ніж просто використання трьох схем навчання і

надання їм можливості голосувати за результат. Навчання всіх трьох схем на відфільтрованих даних і надання їм можливості голосувати може дати ще кращі результати. Однак існує небезпека в методах голосування: Деякі алгоритми навчання краще підходять для певних типів даних, ніж інші, і найбільш підходяща схема може просто не отримати більшості голосів.

Крім того, можна використати більш тонкий метод об'єднання результатів різних класифікаторів, який називається стекуванням.Стекінг – це спосіб об'єднання декількох моделей. Він використовується менш широко, ніж пакування та бустінгу, частково через те, що його важко проаналізувати теоретично, а частково через те, що не існує загальноприйнятого способу реалізації, бо основна ідея може бути інтерпретована в різних варіаціях. На відміну від пакування та бустінгу, стекінг зазвичай не використовується для об'єднання однотипних моделей. Замість цього його застосовують до моделей, побудованих різними алгоритмами навчання. Звичайною процедурою було б оцінити очікувану похибку кожного алгоритму шляхом перехресної перевірки і вибрати найкращий з них для формування моделі для прогнозування на майбутніх даних. Стекінг намагається дізнатися, які класифікатори є надійними, використовуючи інший алгоритм навчання – метанавчання, щоб дізнатися, як найкраще об'єднати вихідні дані базового навчання.

У стекінгу дані розділяються на дві частини: навчальний і тестовий набори. На першому етапі на навчальному наборі застосовуються кілька базових моделей, які навчаються на різних частках даних. Потім, на тестовому наборі, використовуються передбачення цих моделей, які використовуються як вхідні дані для другого рівня моделі, яка навчається на тестових даних. Цей метод дозволяє уникнути перенавчання і забезпечує більш точні передбачення.

Суть будь якого методу завжди, полягає в тому, щоб проаналізувати дані та подивитися на них різними способами. Одна з можливих небезпек, пов'язаних з підходами фільтрації, полягає в тому, що вони можуть просто приносити в жертву екземпляри певного класу (або групи класів) для підвищення точності для решти класів.

Список використаних джерел:

1. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 1-58. https://doi.org/10.1145/1541880.1541882

2. Witten, I. H., Frank, E., & Hall, M. A. (2016). Data mining: Practical machine learning tools and techniques. Morgan Kaufmann.