ВИКОРИСТАННЯ СПЕЦІАЛІЗОВАНИХ БІБЛІОТЕК ПРИ РОЗВ'ЯЗАННІ ЗАДАЧ КЛАСИФІКАЦІЇ І РЕГРЕСІЇ

Новіков А.В., *yuliya.kholodnyak@tsatu.edu.ua*
Таврійський державний агротехнологічний університет імені Дмитра Моторного

В даній роботі пропонуються нові можливості застосування комп'ютерних технологій для розв'язання задач класифікації і регресії. В технології Data Mining задачу класифікації розглядають як задачу визначення значення одного з параметрів аналізованого об'єкту на підставі значень інших параметрів. Параметр, значення якого треба визначити, часто називають залежною змінною, а параметри, що беруть участь в його визначенні, - незалежними змінними.

У розглянутому прикладі незалежними змінними є зарплата, вік, кількість дітей. Залежною змінною в цьому прикладі є кредитоспроможність клієнта. Якщо значеннями незалежних і залежної змінних є дійсні числа, то задача називається задачею регресії. Прикладом задачі регресії може бути задача визначення суми кредиту, яка може бути видана клієнту.

Задачі класифікації і регресії розв'язуються в два етапи. На першому виділяється навчальна вибірка. У неї входять об'єкти, для яких відомі значення як незалежних, так і залежних змінних. У описаному прикладі такою навчальною вибіркою може бути інформація про клієнтів, яким раніше видавалися кредити на різні суми, і інформація про їх повернення.

На підставі навчальної вибірки будується модель дерева рішень для отримання правил класифікації або регресії. Цю модель часто називають функцією класифікації або регресії. Для отримання максимально точної функції до навчальної вибірки пред'являються наступні основні вимоги:

- кількість об'єктів, що входять у вибірку, має бути достатнє великою для більшої точності функції класифікації або регресії;
- у вибірку повинні входити об'єкти, що представляють всі можливі класи в разі задачі класифікації або всю область значень в разі задачі регресії.

На другому етапі побудовану модель дерева рішень застосовують до аналізованих об'єктів для визначення значення залежної змінної.

В результаті побудови моделі дерева рішень отримані наступні правила класифікації:

Rules:

1. *IF вік equals below20 THEN 'покинув' = 'no'*
2. *IF вік equals 20to30 THEN 'покинув' = 'no'*
3. *IF вік equals 31to40 AND поточний_тариф equals normal THEN 'покинув' = 'yes'*
4. *IF вік equals 31to40 AND поточний_тариф equals power THEN 'покинув' = 'no'*
5. *IF вік equals 31to40 AND поточний_тариф equals economy THEN 'покинув' = 'yes'*
6. *IF вік equals 41to50 AND стать equals f THEN 'покинув' = 'no'*
7. *IF вік equals 41to50 AND стать equals m THEN 'покинув' = 'yes'*
8. *IF вік equals 51to60 THEN 'покинув' = 'no'*
9. *IF вік equals above61 THEN 'покинув' = 'no'*

Подібний підхід дозволяє підвищити швидкість аналізу і знизити вимоги до пам'яті завдяки обробці менших обсягів даних в один прохід. Крім того, в цьому випадку аналітичну обробку можна розпаралелити, що позитивно позначається на витраченому часі.

Список використаних джерел

1. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа данных: OLAP и Data Mining. СПб.: БХВ-Петербург, 2004. 336 с.

2. Мандель И.Д. Кластерный аналіз: финансы и статистика, 1988. 176 с.

Науковий керівник: Холодняк Ю.В., к.т.н., ст. викладач