УДК: 004.03

РОЗВ'ЯЗАННЯ ЗАДАЧ КЛАСИФІКАЦІЇ І РЕГРЕСІЇ ІЗ ЗАСТОСУВАННЯМ СПЕЦІАЛІЗОВАНИХ БІБЛІОТЕК

Новіков А.В., Холодняк Ю.В.
Таврійський державний агротехнологічний університет ім. Дмитра Моторного, м. Мелітополь

Пропонуються нові можливості застосування комп'ютерних технологій для розв'язання задач класифікації і регресії.
Ключові слова: *комп'ютерні технології, класифікація, регресія, аналіз даних.*

New opportunities of employment of computer technologies for decision of problems of classification and regression are offered.
Keywords: *computer technology, classification, regression, data analysis.*

З появою перших комп'ютерів настав етап інформатизації різних сторін людської діяльності. В даний час сучасні обчислювальні системи і комп'ютерні мережі дозволяють зберігати великі масиви даних для розв'язання задач обробки і аналізу. На жаль, сама по собі машинна форма представлення даних містить інформацію, необхідну людині, в прихованому вигляді, і для її отримання потрібно використовувати спеціальні методи аналізу даних.

Технологія Data Mining (інтелектуальний аналіз даних) вивчає процес знаходження нових і корисних знань в базах даних. При аналізі накопичених даних часто потрібно визначити, до якого з відомих класів відносяться досліджувані об'єкти, тобто розв'язати задачу класифікації. Наприклад, коли людина звертається в банк за наданням йому кредиту, банківський службовець повинен прийняти рішення: чи кредитоспроможний потенційний клієнт чи ні. Очевидно, що таке рішення приймається на підставі даних про досліджуваний об'єкт (в даному випадку - про людину): його місце роботи, розмір заробітної плати, вік, склад сім'ї і тому подібне. В результаті аналізу цієї інформації банківський службовець повинен віднести людину до одного з двох відомих класів: "кредитоспроможний" і "некредитоспроможний".

В даній роботі пропонуються нові можливості застосування комп'ютерних технологій для розв'язання задач класифікації і регресії**.** У 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 раніше 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 великим, чим більше об'єктів, тим точніше буде побудована на її основі функція 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.

При, наприклад, розв'язанні задачі класифікації даних про клієнтів телекомунікаційної компанії, з метою визначення чи покине клієнт компанію після двох років співпраці, вихідні дані представлені в таблиці 1.

Таблиця 1

Стать	Вік	Поточний тариф	Витрачена сума	Покинув
f	23	Normal	345	No
m	18	Power	9455	No
m	36	Power	456	No
m	34	Normal	3854	Yes
f	52	Economy	2445	No
f	19	Economy	14 326	No
f	45	Normal	347	No
m	42	Economy	5 698	Yes
m	21	Power	267	No
m	48	Normal	4 711	Yes

В результаті побудови моделі дерева рішень отримуємо 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 econoty 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'

Аналізуючи результати, можна зробити висновок, що з 4 аналізованих клієнтів 2 клієнти покинуть компанію.

Висновки. Запропоновано нові можливості для розв'язання задач класифікації і регресії з використанням бібліотеки Xelopes алгоритмів data mining.

Інформаційні джерела

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. Гайдышев И. Анализ и обработка данных: специальный справочник– СПб.: Питер, 2001. – 752 с.

3. Мандель И.Д. Кластерный аналіз : Финансы и статистика, 1988. – 176 с.

4. Мацулевич О.Є., Щербина В.М. Використання пакету прикладних програм NETCRACKER // Фундаментальна підготовка фахівців у природничо-математичній, технічній, агротехнологічній та економічній галузях : матеріали Всеукраїнської наук.-практ. конференції з міжнар. участю (Мелітополь, 11-13 вересня 2017 р.) : присвяченої 85-річчю кафедри вищої математики і фізики ТДАТУ.

5. Мацулевич О.Є., Щербина В.М., Коломієць С.М. Геометричне моделювання складних тривимірних поверхонь із застосуванням матричного рівняння еліптичного повороту // / Праці Таврійського державного агротехнологічного університету, Вип. 19(2), С. 294-300.