УДК 004.89

ОЦІНКА ОСНОВНИХ МЕТОДІВ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ ДЛЯ ЗАДАЧІ ПІДБОРУ ПОКУПОК

Петришин Сергій, Замковий Олександр

Вінницький національний технічний університет

Анотація

Розкрито поняття «інтелектуального аналізу даних». Розглянуто алгоритм дерева прийняття рішень,алгоритм Байеса, алгоритм лінійної регресії. Досліджено переваги та недоліки алгоритмів. Обрано найдоцільніший алгоритм для розв'язання задачі інтелектуального підбору покупок.

Abstract

Disclosed the concept of data mining. The decision tree algorithm, Bayes algorithm, linear regression algorithm are considered. The advantages and disadvantages of algorithms are explored. The most suitable algorithm for the smart purchasing task is chosen.

Вступ

В наш час набирає популярність система «розумний» будинок. Одним зі складових такої системи може бути «розумний» холодильник, який може перевіряти наявність певних продуктів. Завдяки впровадженню системи інтелектуального підбору покупок, «розумний» холодильник може складати список покупок, та замовляти їх через мережу Інтернет. Завдяки цій системі власники «розумного» холодильника завжди будуть мати перелік продуктів, які звикли споживати.

Оцінка основних методів даних для задачі підбору покупок

Основними методами для інтелектуального аналізу даних для задачі підбору покупок є

- дерево прийняття рішень;
- спрощений алгоритм Баєса;
- алгоритм лінійної регресії.

Дерево прийняття рішень — використовується в галузі статистики та аналізу даних для прогнозних моделей. Структура дерева містить такі елементи: «листя» і «гілки». На ребрах («гілках») дерева прийняття рішення записані атрибути, від яких залежить цільова функція, в «листі» записані значення цільової функції, а в інших вузлах — атрибути, за якими розрізняються випадки[1]. Щоб класифікувати новий випадок, треба спуститися по дереву до листа і видати відповідне значення. Подібні дерева рішень широко використовуються в інтелектуальному аналізі даних для задачі підбору покупок. Мета полягає в тому, щоб створити модель, яка прогнозує значення цільової змінної на основі декількох змінних на вході. Кожен лист являє собою значення цільової змінної, зміненої в ході руху від кореня по листа. Кожен внутрішній вузол відповідає одній з вхідних змінних. Дерево може бути також «вивчено» поділом вихідних наборів змінних на підмножини, що засновані на тестуванні значень атрибутів. Це процес, який повторюється на кожній з отриманих підмножин. Рекурсія завершується тоді, коли підмножина в вузлі має ті ж значення цільової змінної, таким чином, воно не додає цінності для прогнозування. Процес, що йде «згори донизу», індукція дерев рішень (TDIDT), є прикладом поглинаючого «жадібного» алгоритму, і на сьогодні є найбільш поширеною стратегією дерев рішень для даних, але це не єдина можлива стратегія. В інтелектуальному аналізі даних, для розв'язання задачі підбору покупок, дерева рішень можуть бути використані як математичні та обчислювальні методи, щоб допомогти

описати, класифікувати і узагальнити набір даних, які можуть бути записані таким чином:

$$(x, Y) = (x_1, x_2, x_3 \dots x_k, Y)$$

Залежна змінна Y є цільовою змінною, яку необхідно проаналізувати, класифікувати й узагальнити. Вектор x складається з вхідних змінних x_1, x_2, x_3 тощо, які використовуються для виконання цього завдання[1].

Спрощений алгоритм Байеса є алгоритмом класифікації, на підставі теореми Байеса і використовується в прогнозуючому моделюванні. Слово «спрощений» в його назві вказує на те, що алгоритм використовує методи Байеса, але не враховує можливі залежності. Даний алгоритм вимагає меншої кількості обчислень, ніж інші алгоритми, і може бути використаним для швидкого формування моделей інтелектуального аналізу даних для виявлення відносин між вхідними і прогнозованими стовпцями, також цей алгоритм підходить для розв'язання задачі підбору покупок.

Абстрактно імовірнісна модель для класифікатора — це умовна модель

$$p(C \mid F_1,\dots,F_n)$$

над залежною змінною класу C з малою кількістю результатів або класів, залежна від кількох змінних $F_1,\dots,F_n$[2]. Проблема полягає в тому, що коли кількість властивостей n, тобто кількість можливих комбінацій покупок, дуже велика або коли властивість може приймати велику кількість значень, тоді будувати таку модель на імовірнісних таблицях стає неможливо. Тому ми переформулюємо модель, щоб зробити її такою, яка легко піддається обробці.

Використовуючи теорему Баєса, запишемо

$$p(C \mid F_1,\dots,F_n)=p(C)p(F_1,\dots,F_n)|C)/p(F_1,\dots,F_n) .$$

На практиці цікавий лише чисельник цього дробу, так як знаменник не залежить від C і значення властивостей дані, так що знаменник — константа[2]. Цей алгоритм можна використовувати для початкового дослідження даних, а потім застосувати результати для створення додаткових моделей інтелектуального аналізу з іншими алгоритмами, які вимагають більшої кількості обчислень і є більш точними, що в кінцевому рахунку приведе до розв'язання задачі підбору покупок.

Алгоритм лінійної регресії є різновидом алгоритму дерева прийняття рішень, що допомагає розрахувати лінійну зв'язок між залежною і незалежною змінною, а потім використовувати цей зв'язок при прогнозуванні, або ж підборі покупок. При виборі алгоритму лінійної регресії викликається особливий варіант алгоритму дерева прийняття рішень з параметрами, які обмежують поведінку алгоритму і вимагають використання певних типів даних на вході. Більш того, в моделі лінійної регресії для обчислення зв'язків при початковому проході використовується весь набір даних; тоді як в стандартній моделі дерева прийняття рішення дані багаторазово розбиваються на менші підмножини або дерева[3].

У ході оцінювання методів інтелектуального аналізу даних для задачі підбору покупок було розглянуто такі методи, як дерево прийняття рішень, спрощений алгоритм Баєса, алгоритм лінійної регресії. Оскільки покупки зазвичай пов'язані між собою, то для задачі підбору покупок доцільно використовувати алгоритм дерева прийняття рішень в комбінації з спрощеним алгоритмом Баєса, оскільки це забезпечить можливість перегляду значення цільової функції з додаванням нових параметрів, та зміною уже існуючих параметрів.

Список використаних джерел:

1.Дерево прийняття рішень - http://stud.com.ua/31896/menedzhment.

2. Левитин А. Алгоритмы. Введение в разработку и анализ. Вильямс, 2006. Р. 160.

3. Алгоритм лінійної регресії – [Електронний ресурс]. – Режим доступу: https://msdn.microsoft.com/ru-ru/library/ms174824(v=sql.120).aspx