УДК 004.04

О. Д. Замковий
С. І. Петришин

ОЦІНКА АКТУАЛЬНОСТЕЙ МЕТОДІВ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ ДЛЯ ЗАДАЧІ ПІДБОРУ ПОКУПОК

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 задачі інтелектуального підбору покупок.

Ключові слова: інтелектуальний аналіз даних, статистика, дані, дерево прийняття рішень, спрощений алгоритм Байеса, алгоритм лінійної регресії.

Abstract

Disclosed the concept of data mining. The decision tree 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..

Keywords: data maning, statistics, data, decision tree, simplified Bayesian algorithm, linear regression algorithm.

В наш час набирає популярність система «розумний» будинок. Одним зі складових такої системи може бути «розумний» холодильник, який може «моніторити» 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.

Дані — це інформація (найчастіше цифрова), подана у формалізованому вигляді, прийнятному для обробки автоматичними засобами за можливої участі людини. Дані - інформація, одержана в експерименті, взята з опублікованих праць чи отримана в результаті розрахунків. При цьому мають бути точно описані умови їх отримання та способи розрахунків. Представляються та організовуються у спосіб зручний для подальшої обробки та аналізу.

Статистика — наука, що вивчає методи кількісного охоплення і дослідження масових, зокрема суспільних, явищ і процесів. А також власне кількісний облік масових явищ. Зокрема, облік у будь-якій галузі господарства, суспільного життя, що здійснюється методами цієї науки, а також дані цього обліку. Статистика вивчає кількісний бік масових явищ і процесів у нерозривному зв'язку з їх якісним боком. Статистика поділяється на математичну та прикладну. Прикладну статистику поділяють за галузями на демографічну, економічну, фінансову, соціальну, санітарну, судову, біологічну, технічну тощо.

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. Щоб класифікувати новий випадок, треба спуститися по дереву до листа і видати відповідне значення. Подібні дерева рішень широко використовуються в інтелектуальному аналізі даних. 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, 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. 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[2].

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.

Алгоритм лінійної регресії є різновидом алгоритму дерева прийняття рішень. 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].

Оскільки покупки зазвичай пов’язані між собою, то доцільно використовувати алгоритм дерева прийняття рішень, оскільки він забезпечує можливість перегляду значення цільової функції з додаванням нових параметрів, та зміною уже існуючих параметрів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Сергей Николенко, Александр Тулупьев. Самообучающиеся системы. Москва, 2009. P. 288.
2. Левитин А. Алгоритмы. Введение в разработку и анализ. Вильямс, 2006. P. 160.
3. Алгоритм лінійної регресії – [Електронний ресурс]. – Режим доступу: https://msdn.microsoft.com/ru-ru/library/ms174824(v=sql.120).aspx

Замковий Олександр Дмитрович — студент групи 1КН-14б, факультет інформаційних технологій та комп’ютерної інженерії, Вінницький національний технічний університет, м. Вінниця.

Петришин Сергій Іванович, – асистент кафедри комп'ютерних наук ВНТУ, Вінницький національний технічний університет, м. Вінниця.

Zamkovti Oleksandr D. — Faculty of Information Technologies and Computer Engineering, Vinnytsia National Technical University, Vinnytsia.

Sergiy I. Petrishyn — assistant of the Computer Sciences Chair, Vinnytsia National Technical University, Vinnytsia.