УДК

Д. В. Кукуруза
С. І. Петришин

ДОСЛІДЖЕННЯ ОСНОВНИХ 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.

Ключові слова:

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.

Моделювання - метод пізнання навколишнього світу, який можна віднести до загальнонаукових методів, що застосовуються як на емпіричному, так і на теоретичному рівні пізнання. При побудові та дослідженні моделі можуть застосовуватися практично всі інші методи пізнання.

Під моделлю розуміється такий матеріальний чи уявно представлений об'єкт, який в процесі пізнання (вивчення) заміщає об'єкт-оригінал, зберігаючи деякі важливі для даного дослідження типові його риси.

Наприклад, існує завдання прогнозування попиту. Для задачі прогнозування попиту товарів полягає в тому, що необхідно передбачити, які продукти необхідно закупити та у якій кількості. Якщо закупити більше, ніж потрібно, то, якщо товар швидко псується, він зіпсується, в іншому випадку він буде займати місце на складі, за оренду якого теж необхідно платити. Якщо закупити занадто мало товару, то на сайті весь час буде показано, що його немає в наявності, і користувачі здійснюватимуть покупки у конкурентів. Тобто необхідно передбачати кількість товару, яке буде затребуваною. Тут виникає питання про раціональну закупку товарів.

Вивчаючи методи прийняття рішень необхідно брати до уваги два нюанси. По-перше, приймати рішення не так складно як здається, але прийняти справді правильне рішення дійсно важко. По-друге, прийняття рішень не завжди піддається логіці – інколи ним керують почуття.

По цій причині рішення можуть бути як спонтанними й нелогічними, як логічними та обдуманими.

Розглянемо методи, які найчастіше використовуються в області статистики й аналізу даних для прогнозних моделей [1].

На початку 80-х років в дослідженнях зі штучного інтелекту сформувався самостійний напрямок, який одержав назву "експертні системи" (ЕС). Основним призначенням ЕС є розробка програмних засобів, які при вирішенні завдань, важких для людини, одержують результати, які не поступаються за якістю і ефективністю рішення, рішенням одержуваних людиною-експертом. ЕС використовуються для вирішення так званих неформалізованих задач, загальним для яких є те, що:

- задачі не можуть бути задані в числовій формі;
- мету не можна виразити в термінах точно визначеної цільової функції;
- не існує алгоритмічного рішення задачі;

- якщо алгоритмічне рішення є, то його не можна використовувати через обмеженість ресурсів (час, пам'ять).

Крім того неформалізовані задачі характеризуються помилковістю, неповнотою, неоднозначністю і суперечливістю як вихідних даних, так і знань про розв'язуваній задачі.

Недоліки експертних систем:

- творчий потенціал: людина-експерт може реагувати творчо на незвичайні ситуації, експертні системи не можуть;
- навчання: людина-експерт автоматично адаптуються до зміни середовища; експертні системи потрібно явно модифікувати;
- експертні системи не хороші, якщо рішення не існує або коли проблема лежить поза області їх компетенції.

Клас експертних систем сьогодні об'єднує кілька тисяч різних програмних комплексів, які можна класифікувати за різними критеріями: вирішити завдання, зв'язок з реальним часом, тип ЕОМ, ступінь інтеграції.

Алгоритм лінійної регресії є різновидом алгоритму дерева прийняття рішень, що допомагає розрахувати лінійний 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[2].

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[3].

Байєсівські методи розроблені внаслідок численних спроб вчених визначити проблеми статистичного аналізу поведінки різних процесів і знайти їх рішення за допомогою застосування основи байєсівської методології - теореми Байєса. Використання даної теореми має ряд передумов, основна з яких - наявність певних співвідношень між вірогідністю явищ, що мають різний характер і специфікації будь-якого явища на потрібному рівні.

Байєсівськаа методологія відрізняється від інших підходів тим, що ще до отримання даних дослідник визначає рівень своєї довіри до можливих моделей і згодом представляє її у вигляді певних ймовірностей.

Необхідно виділити наступні особливості байєсівського підходу:

- абсолютно всі параметри і величини прийнято вважати випадковими, а саме точне значення параметрів невідомо досліднику, з чого випливає те, що параметри є випадковими з точки зору незнання дослідника;

методи Байеса використовуються навіть при нульовому обсязі вибірки [4].

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. Ю. А. Зав'ялець Комп'ютерні мережі / Ю. А. Зав'ялець - Чернівці – 2006. 182 с.
2. Сергей Николенко, Александр Тулупьев. Самообучающиеся системы. Москва, 2009. P. 288.
3. Левитин А. Алгоритмы. Введение в разработку и анализ. Вильямс, 2006. P. 160.
4. Алгоритм лінійної регресії – [Електронний ресурс]. – Режим доступу: https://msdn.microsoft.com/ru-ru/library/ms174824(v=sql.120).aspx

Петришин Сергій Іванович, *старший викладач кафедри комп'ютерних наук факультету інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, Вінниця, petryshyn@vntu.edu.ua.*

Кукуруза Дмитро Вячеславович, *1КН-18мс, факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, Вінниця.*

Sergiy Petryshyn*, Senior Lecturer, Computer Science, Faculty of Information Technology and Computer Engineering, Vinnitsa National Technical University, Vinnytsia, petryshyn@vntu.edu.ua.*

Kukuruza Dmitry*, 1KN-18ms, Faculty of Information Technology and Computer Engineering, Vinnitsa National Technical University, Vinnytsia.*