

ВИКОРИСТАННЯ ДЕРЕВ РІШЕНЬ ДЛЯ ВИЗНАЧЕННЯ ПОТЕНЦІЙНИХ ПОКУПЦІВ ТОВАРУ

Вінницький національний технічний університет

Анотація

Розглянуто питання вибору інструментальних запитів для визначення потенційних покупців товару. Проаналізовано випадки доцільності використання дерев рішень для пошуку асоціативних правил.

Ключові слова: інтелектуальний аналіз даних, дерево рішень, асоціативні правила, ентропія

Abstract

The issues of the selection of tool requests to determine potential buyers of the goods are considered. Cases the feasibility of using decision trees for mining Association rules are analyzed.

Keywords: Data mining, decision tree, association rules, entropy

Однією з найважливіших задач маркетингу є моделювання поведінки покупця [1]. Зокрема, важливо правильно здійснити сегментацію покупців, рекламні заходи, визначення так званих «сукупних» товарів, які часто купуються разом або ніколи разом не купуються.

Одним з найбільш розповсюджених методів Data Mining для розв'язання останньої задачі є афінитивний аналіз (affinity analysis), метою якого є дослідження зв'язків між подіями, що відбуваються сумісно, і виявлення правил (асоціацій), для їх кількісного опису. Базовим поняттям теорії асоціативних правил є транзакція – деяка множина подій, що відбуваються сумісно. Наприклад, купівельні транзакції є набором товарів, куплених покупцем за один візит.

Отож алгоритм асоціативних правил призначений для виявлення в даних залежностей типу $A \rightarrow B$, де A і B є наборами пар атрибут = значення [2].

Правило $A \rightarrow B$ характеризується підтримкою (support), значення якої S співпадає з імовірністю $P(A \& B)$ сумісної появи в одній транзакції товарів A і B :

$$S(A \rightarrow B) = P(A \& B) = D_{A \& B} / D, \quad (1)$$

де D - загальна кількість транзакцій

$D_{A \& B}$ - кількість транзакцій, що містить товари A і B .

Мірою точності правила $A \rightarrow B$ є його достовірність (confidence), яка співпадає з умовною імовірністю купівлі товару B за умов покупки товару A :

$$C(A \rightarrow B) = P(B|A) = P(A \& B) / P(A) = \frac{D_{A \& B}}{D} / \frac{D_A}{D} = D_{A \& B} / D_A, \quad (2)$$

де D_A - кількість транзакцій, що містить товар A .

Алгоритми пошуку асоціативних правил мають знаходити всі правила $A \rightarrow B$, підтримка і достовірність яких є вищими за деякі наперед визначені мінімальні пороги.

Разом з тим, Data Mining має й інші інструментальні засоби для виявлення закономірностей, зокрема такі, як Байєсовські класифікатори і дерева рішень. Метою даної роботи було визначення доцільності використання дерев рішень для розв'язання задач виявлення асоціацій

Алгоритм дерева рішень призначений для вирішення задач регресії та класифікації, тобто для виявлення залежності цільового параметра від значення інших параметрів. Процедура побудови дерева рішень по навчальній множині T об'єктів (записів, прикладів) заснована на ітераційному обчисленні ступеня впливу кожного вхідного атрибуту на значення цільового атрибуту з точки зору приросту інформації і використовує атрибут, що впливає на цільовий атрибут найбільшою мірою для розбиття вузла дерева рішень [3]:

$$Gain(X) = Info(T) - Info_x(T), \quad (3)$$

де $Info(T)$ – ентропія множини T , а

$$Info_x(T) = \sum_{i=1}^n \frac{T_i}{T} \times Info_x(T_i), \quad (4)$$

де кожна підмножина T_i містить усі приклади зі значенням O_i обраного ознаки O .

Вузол верхнього рівня описує розподіл значень вихідного атрибута по всій сукупності даних. Кожен наступний вузол описується розподілом вихідного атрибута при дотриманні умов на значення вхідних атрибутів у цьому вузлі. Дерево росте доки, поки дочірні вузли збільшують якість прогнозу. Зростання припиняється також коли кількість записів в базі даних, що описують вузол, стає меншою за визначений рівень. Алгоритм здійснює пошук атрибутів і їх значень, розбиття по яким дозволяє з найбільшою ймовірністю правильно передбачити значення вихідного атрибута.

Незважаючи на схожість розв'язуваної задачі і форми подання результатів, правила, одержувані обома алгоритмами, в окремих випадках можуть значно відрізнятись, оскільки процес побудови дерева рішень заснований на максимізації приросту інформації, в той час як алгоритм асоціативних правил заснований на виділенні «частих» наборів з високою частотою появи однієї частини (права частина правила) щодо наборів, в яких є інша частина (ліва частина правила).

Проведені дослідження доводять доцільність використання для виявлення асоціативних правил дерев рішень, зокрема у випадках правил з малою підтримкою.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Процесс принятия решения о покупке [Электронный ресурс] / Энциклопедия экономиста. – Режим доступа: <http://www.grandars.ru/student/marketing/process-prinyatiya-resheniya-o-pokupke.html>
2. Паклин Н. Б., Орешков .В. И. Бизнес-аналитика: от данных к знаниям (+CD): Учебное пособие, 2-е изд., испр. – СПб.: Питер.ю 2013. – 704 с. – ISBN 978-5-459-00717-6.
3. Барсегян А.А. Анализ данных и процессов: учеб. Пособие / А. А. Барсегян, М. С. Куприянов, И. И. Холод, М. Д. Тесс . С. И. Елизаров – 3-е изд., перераб. и доп. СПб: БХВ-Петербург, 2009. – 512 с. – ISBN 978-5-9775-0368-6.

Оленчук Дмитро Сергійович – студент групи ІКН-14мс, факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, м. Вінниця, e-mail: ihaveadreams@mail.ru.

Месюра Володимир Іванович – к.т.н., доцент, професор кафедри комп'ютерних наук, Вінницький національний технічний університет, м. Вінниця.

Dmytro S. Olenchuk – Student of Department of Information Technology and Computer Engineering, Vinnytsia National Technical University, Vinnytsia, e-mail: ihaveadreams@mail.ru.

Volodymyr I. Mesyura – Cand. Sc. (Eng), Assistant Professor, Professor of the Computer Science Chair, Vinnytsia National Technical University, Vinnytsia.