

ДОСЛІДЖЕННЯ МЕТОДІВ CRM НА БАЗІ ГЕНЕРАЦІЇ АСОЦІАТИВНИХ ПРАВИЛ ТА ЇХ ВИКОРИСТАННЯ В ІНФОРМАЦІЙНІЙ СИСТЕМІ БІБЛІОТЕКИ

Забийворота М.А., Тітов С.В

Кафедра системотехніки
Харківський національний університет
радіоелектроніки, Україна

E-mail: mykhailo.zabyivorota@nure.ua,

serhii.titov@nure.ua

Abstract

The text explains the concept and application of association rule learning, a data mining and machine learning method used to find meaningful relationships between variables in large datasets. In this article I define key measures used in association rule learning — Support, Confidence, and Lift — and explain their roles in evaluating the strength and reliability of discovered patterns. For instance, Support measures how often items appear together, Confidence estimates the likelihood that one item appears given the presence of another, and Lift evaluates how much one item's presence increases the probability of the other appearing. Based on this knowledge I explain the advantages and disadvantages of using association rules in the design of an information system, particularly a library recommendation system. Advantages include simplicity, interpretability, efficiency on large datasets, and the lack of need for personal user data. Disadvantages involve the cold start problem, the inability to consider sequence or context, and the generation of too many trivial rules.

Як люди, ми природньо схильні встановлювати зв'язки між різними спостереженнями та ідеями. Ці зв'язки ґрунтуються на асоціаціях, із якими ми себе ототожнюємо.

Ми схильні швидко робити висновки, плутаючи асоціацію з причинно-наслідковим зв'язком або кореляцією. У навчанні асоціативних правил усе інакше. Спершу розглянемо, що це таке і чому воно не обов'язково означає наявність причинності.

Коли ми заходимо до супермаркету, є причина, чому пиво розташовано поруч з чіпсами та іншими товарами «до пива». Таке розміщення товарів – результат аналізу покупок минулих клієнтів. Його мета – зробити пошук потрібних товарів зручним і водночас спонукати купити більше, ніж планував відвідувач. Він міг зайти тільки за пивом, але, побачивши чіпси, підсвідомо відчув зв'язок і придбав їх.

Навчання асоціативних правил – це метод машинного навчання, який базується на правилах і використовується для виявлення взаємозв'язків між змінними у великих базах даних. Його мета – знаходити сильні, статично значущі зв'язки між об'єктами, використовуючи певні міри «цікавості» таких зв'язків.

Асоціативні правила це один з методів інтелектуального аналізу даних, який дозволяє знаходити закономірності та зв'язки між об'єктами у великих наборах даних. Ці правила представляються у вигляді «Якщо {A}, то {B}». У контексті бібліотеки це означає: якщо читач взяв книгу (або набір книг) A, то з високою ймовірністю він також візьме книгу (або набір книг) B.

Наприклад, читачі, які беруть «Дюну», часто цікавляться і «Гіперіоном».

Але чи означає це що покупець завжди купуватиме чіпси, якщо він придбав пиво. Або читач "Дюни" зацікавиться "Гіперіоном".

Не обов'язково. Навчання асоціативних правил лише кількісно оцінює силу спільної появи двох або більше елементів. Наприклад, ви можете захотіти купити новий телефон, якщо ваш друг це зробив, але це не означає що ви точно зробите це. Так само метод не враховує порядок дій – він розглядає всі елементи покупки як єдину множину без послідовності. Через це дуже складно сказати, чому саме

клієнт надає перевагу. Не всі закономірності однакової сили, і важливо вміти розрізняти, де просто збіг, а де – справді значущий зв'язок.

Як це перевірити? Припустимо, що існують різні набори подій, пов'язані з реакцією мого kota на те, що його зачиняють в кімнаті. Ось можливі набори:

Підтримка (Support): Співпадіння ситуацій, коли кіт був зачинений у кімнаті та сердитий, траплялося значно частіше, ніж коли він був зачинений та спокійний. Кількість випадків, коли він злився, безумовно перевищує кількість разів, коли він поведився спокійно. Крім того, злість майже завжди супроводжується шипінням {Шипіння}, тоді як спокій — сном {Сон}. Іншими словами, набір подій №3 має більшу підтримку, ніж набір №4.

Якщо відомо, що після певного порогу кіт стабільно поводить певним чином, то можна вважати це значущою підтримкою. Однак якщо рівень підтримки низький, то й сила асоціації буде слабкою — не можна зробити надійних висновків у такій ситуації.

Довіра (Confidence): Бажано мати впевненість у зв'язках у нашому житті. Якщо в наборі подій є {Шипіння}, то скільки разів при цьому кіт справді був зачинений у кімнаті {Зачинений}?

Ця умовна ймовірність показує, наскільки впевнено можна стверджувати, що певне правило асоціації веде до іншого. Іноді це значення може суперечити інтуїтивному розумінню. Якщо кількість спостережень загалом невелика, але спостерігається значне перекриття між різними подіями, рівень довіри може бути високим – навіть близьким до 1. Хоча інтуїтивно ми розуміємо, що такий зв'язок може бути хибним. Наприклад, якщо дано {Сон}, то який рівень довіри для {Зачинений}. Саме тому потрібна ще одна міра, щоб не порівнювати червоне з гарячим.

Підйом (Lift): Підйом визначає, наскільки підтримка одного набору залежить від іншого. Або інакше — який підйом має довіра до {Сон} у контексті наявності {Зачинений} у тому ж наборі. Ми аналізуємо, наскільки змінюється ймовірність появи {Зачинений} у наборі, якщо відомо, що там також є {Сон}, порівняно з випадком, коли ми нічого не знаємо про існування {Сон}.

Коли наявність одного елемента підвищує ймовірність появи іншого в наборі, підйом буде більшим за 1. Це свідчить про сильну асоціацію, часту спільну появу та більшу надійність правила $\{X\} \rightarrow \{Y\}$. Саме ця метрика допомагає у випадках, подібних до розміщення товарів у супермаркеті, про що ми говорили раніше — вона дозволяє визначити, які товари дійсно пов'язані між собою і розташування яких поруч має сенс.

$$\begin{aligned}
 \text{Rule } X \Rightarrow Y & \begin{cases} \text{Support} = \frac{\text{Frequency}(X,Y)}{N} \\ \text{Confidence} = \frac{\text{Frequency}(X,Y)}{\text{Frequency}(X)} \\ \text{Lift} = \frac{\text{Support}}{\text{Support}(X) * \text{Support}(Y)} \end{cases}
 \end{aligned}$$

Рис. 1. Вплив Підтримки, Довіри та Підйому на правило

При розробці інформаційної системи метод асоціативних правил має деякі переваги та недоліки над іншими рекомендаційними алгоритмами. Нижче наведено переваги та недоліки при використанні методу асоціативних правил для проектування інформаційної системи бібліотеки.

Переваги:

1. Простота та інтерпретованість. Правила є зрозумілими для людини. Їх легко пояснити та інтерпретувати.
2. Не вимагає персональних даних. Для побудови правил не потрібні оцінки чи історія транзакцій конкретного користувача, лише дані про видачі книг.

3. Ефективність на великих даних. Існують добре відомі та оптимізовані алгоритми для знаходження правил у баз даних. Наприклад, Apriori.

Недоліки:

1. Холодний старт. Алгоритм не може давати рекомендації для нових книг, оскільки для них ще нема про спільне взяття з іншими

2. Не враховує послідовність. Асоціативні правила не беруть до уваги порядок, у якому читач брав книги, що може бути важливим для серій книг або певних тем.

3. Велика кількість правил. Може генеруватися велика кількість правил, більшість з яких є тривіальними.

4. Не враховує контекст. Наприклад, сезонність інтересу до певних тем або мету читання (розвага, навчання).

Розробка алгоритму асоціативних правил була реалізована за використання мови програмування Python та інтегрування його до інформаційної системи, що написана на Java. Під час розробки інформаційного забезпечення був проведений обґрунтований аналіз вибору СУБД та спроектована база даних.

Під час аналізу предметної області була визначена сфера застосування CRM систем в інформаційних центрах (бібліотеках), проведено аналіз вже існуючих інформаційних систем, а також визначено специфічні особливості для предметної області, що розглядається.

Для визначення того, який підхід найбільше підходить для розробки методу CRM, було проведено аналіз та дослідження наступних існуючих методів побудови рекомендацій: фільтрації на основі контексту/вмісту, колаборативна фільтрація, гібридна фільтрація та метод генерації асоціативних правил. В результаті дослідження було визначено, що використання методу генерації асоціативних правил є найбільш доцільним для специфіки розглянутої предметної області.

Для розробки методів CRM системи з урахуванням специфіки предметної області був застосований алгоритм асоціативних правил Apriori. Розроблений метод опирається на особливості предметної області. Він враховує специфіку бібліотечного фонду та уникає рекомендацій з несумісними типами та жанрами літератури. Після розробки методу були описані основні бізнес-процеси та визначені функціональні вимоги до інформаційної системи.

Література

1. Agrawal R., Imieliński T., Swami A. Mining association rules between sets of items in large databases // Proceedings of the 1993 ACM SIGMOD international conference on Management of data. Washington, D.C., 1993. P. 207–216.
2. Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques. 3rd ed. Morgan Kaufmann Publishers, 2011. 744 p.
3. Tan P.-N., Steinbach M., Kumar V. Introduction to Data Mining. 2nd ed. Pearson, 2018. 769 p.
4. Zaki M. J., Meira Jr. W. Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press, 2014. 595 p.
5. Aggarwal C. C. Data Mining: The Textbook. Springer, 2015. 734 p.
6. Тітов С. В., Тітова О. В., Чорна О. С. Метод знаходження апроксимацій приблизних множин з використанням систем числення. Системи обробки інформації. 2023. № 2 (173). С. 58-62. <https://doi.org/10.30748/soi.2023.173.07>
7. Ситніков Д.Е., Ситнікова П.Е., Тітов С.В., Тітова О.В. Фільтрація результуючого набору асоціативних правил з точки зору оцінки цікавості / Д.Е. Ситніков, П.Е. Ситнікова, С.В. Тітов, О.В. Тітова // Системи обробки інформації № 1(164) 2021, стр. 83-88. <https://doi.org/10.30748/soi.2021.164.09>