Науковий вісник НЛТУ України
Scientific Bulletin of UNFU
http://nv.nltu.edu.ua
https://doi.org/10.15421/40270923
Article received 13.11.2017 p.
Article accepted 28.11.2017 p.
УДК 004.62

ISSN 1994-7836 (print)
ISSN 2519-2477 (online)

@ Correspondence author
I. J. Zheliznyak
iryna.zhelizniak@gmail.com

І. Й. Желізняк
Національний університет "Львівська політехніка", м. Львів, Україна

ПРАВИЛА ПОБУДОВИ АСОЦІАТИВНИХ ПРАВИЛ НА ПРИКЛАДІ ФІЗИЧНИХ ПОКАЗНИКІВ ПАЦІЄНТА

Охарактеризовано правила побудови асоціативних правил. Позначено об'єкти, що складають досліджуваний набір. Побудовано асоціативні правила для призначення аналізів пацієнту. Наведено найпоширеніші терміни інтелектуального аналізу даних. Розглянуто множину транзакцій, які доступні для медичного аналізу пацієнта. Описано послідовність об'єктів та задане відношення порядку. Зазначено мінімальне значення підтримки множини та відсіяно асоціативні правила на основі цього значення. Охарактеризовано величини корисності асоціативних правил, за допомогою яких встановлюється важливість того чи іншого асоціативного правила. Виявлено, що правильна оцінка корисності асоціативного правила впливає на об'єм та швидкодію доступу до інформації. Введено унікальний ідентифікатор для досліджуваного набору аналізів пацієнта. Означено додаткові чисельні атрибути досліджуваних об'єктів. Охарактеризовано транзакції, що містять додаткові атрибути та операції не лише наявності, а також порівняння. Встановлено відмінність між асоціативними правилами та секвенційним аналізом. Оцінено вплив побудови асоціативних правил під час дослідження предметної області та інтелектуального аналізу даних. Отримані результати буде використано у подальших дослідженнях у цій предметній області.

Ключові слова: секвенційний аналіз; data mining; медицина; біоінформатика; виявлення закономірностей.

Вступ. У медичних і біологічних дослідженнях, так само як і в практичній медицині, спектр вирішуваних завдань настільки широкий, що можливе використання будь-яких методологій Data Mining. Прикладом може слугувати побудова діагностичної системи або дослідження ефективності хірургічного втручання (Brin, & Page, 2012; Negnivitsky, 2002).

Одним з найпередовіших напрямків медицини є біоінформатика. Об'єктом дослідження біоінформатики є величезні обсяги інформації про послідовності ДНК і первинну структуру білків, що з'явилися внаслідок вивчення структури геномів мікроорганізмів, ссавців і людини. Абстрагуючись від конкретного змісту цієї інформації, її можна розглядати як набір генетичних текстів, що складаються із протяжних символьних послідовностей. Виявлення структурних закономірностей у таких послідовностях входить до переліку завдань, що ефективно вирішуються засобами Data Mining, наприклад, за допомогою сіквенціального та асоціативного аналізу (ohannes, Setnes & Abonyic, 2001; Sutton & Barto, 2008).

Мета дослідження – виявити найважливіші правила побудови асоціативних правил; визначити закономірності побудови асоціативних правил та поділ фізичних показників на різні рівні ієрархії.

Об'єкти та методика дослідження. Однією з найпоширеніших задач аналізу даних є визначення наборів об'єктів, що часто трапляються у великому наборі об'єктів. Опишемо цю задачу в узагальненому вигляді. Для цього позначимо об'єкти, що складають досліджувані набори (itemsets), так: $I = \{i_1, i_2, \ldots, i_j, \ldots, i_n\}$, де: j_i – об'єкти, що входять у досліджувані набори; n – загальна кількість об'єктів (Negnivitsky, 2002; Zhang, 2001). У сфері медицини такими об'єктами, наприклад, є показники та аналізи пацієнта (табл. 1)

Табл. 1. Об'єкти, що входять у досліджуваний набір

Ідентифікатор	Показник	Значення
0	Артеріальний тиск	120/80 мм рт. ст.
1	Венозний тиск	70 мм вод. ст.
2	Капілярний тиск	70 мм рт. ст.
3	Пульс	85 ударів/хв
4	Температура	36,6 С
5	Рівень гемоглобіну в крові	145 г/л
6	pH	7,35

Отже, вони відповідають такій множині об'єктів:
I = {артеріальний тиск, венозний тиск, капілярний тиск, пульс, температура, рівень гемоглобіну в крові, pH}.

Набори об'єктів із множини I, що зберігаються в БД і піддаються аналізу, називають транзакціями. Опишемо транзакцію як підмножину множини I, а саме:

$$T = \{i_j | i_i \in I\}.$$

Такі транзакції в лікарні відповідають здачі медичних аналізів пацієнта і зберігаються в базі даних у вигляді медичної картки. У них перераховуються аналізи, що пацієнт здав для анамнезу та поставлення діагнозу (Johannes, Setnes, & Abonyic, 2001; Zhang, 2001).

Набір транзакцій, інформація про які доступна для аналізу, опишемо такою множиною:

$$D = \{T_1, T_2, \ldots, T_r, \ldots, T_m\},$$

де m – кількість доступних для аналізу транзакцій.

Інформація про авторів:

Желізняк Ірина Йосифівна, аспірант кафедри інформаційних систем та мереж. **Email:** iryna.zhelizniak@gmail.com

Цитування за ДСТУ: Желізняк І. Й. Правила побудови асоціативних правил на прикладі фізичних показників пацієнта. Науковий вісник НЛТУ України. 2017. Вип. 27(9). С. 107–110.

Citation APA: Zheliznyak, I. J. (2017). Some Regulations for Constructing Associative Rules on the Example of Patient's Physical Characteristics. *Scientific Bulletin of UNFU, 27*(9), 107–110. https://doi.org/10.15421/40270923

Результати дослідження та їх обговорення. Для використання методів Data Mining множина D може бути представлена у вигляді таблиці (табл. 2):

Табл. 2. Множина досліджуваних об'єктів

Номер транзакції	Номер показника	Показник	Значення
0	0	Артеріальний тиск	110/75 мм рт. ст.
0	3	Пульс	110 ударів/хв
0	1	Венозний тиск	58 мм рт. ст.
1	4	Температура	37,4 С
1	5	pH	7,46
2	1	Венозний тиск	72 мм рт. ст.
2	6	pH	7,81
2	4	Температура	37,2 С

Множину транзакцій, у яку входять об'єкти ji, позначають так: $D = \{T_r | i_j \in T_r; j = 1..n; r = 1..m\} \subseteq D$.

У цьому прикладі множиною транзакцій, що містять об'єкт Температура, є така множина:

$D_{\text{температура}}$ = {{Температура, Рівень гемоглобіну}, {Венозний тиск, pH, Температура}}.

Деякий довільний набір об'єктів (itemset) позначимо в такий спосіб: $F = \{i_j | i_j \in I; j = 1..n\}$.

Множини транзакцій, у які входить набір F, позначимо в такий спосіб:

$$DF = \{T_r | F \subseteq T_r; r = 1..m\} \subseteq D.$$

Відношення кількості транзакцій, в яке входить набір F, до загальної кількості транзакцій називають підтримкою (support) набору F і позначають $Supp(F)$

$$Supp(F) = |D_F|/D.$$

Наприклад, для набору {pH, температура} підтримка буде дорівнювати 2/3, тому що цей набір входить у дві транзакції (номери 1 та 2) із трьох можливих (Negnivitsky, 2002; Sutton & Barto, 2008).

Під час пошуку аналітик може вказати мінімальне значення підтримки цікавих йому наборів Suppmin. Набір називають великим (large itemset), якщо значення його підтримки більше від мінімального значення підтримки, заданого користувачем: $Supp(F) > Supp_{min}$.

Отже, під час пошуку асоціативних правил потрібно знайти множину всіх частих наборів

$$L = \{F | Supp(F) > Supp_{min}\}.$$

У цьому випадку наборами при $Supp_{min} = 2/3$, є такі:

- {pH} $Supp_{min} = 2/3$;
- {венозний тиск} $Supp_{min} = 2/3$;
- {температура} $Supp_{min} = 2/3$;
- {pH, температура} $Supp_{min} = 2/3$.

Під час аналізу часто викликає інтерес послідовність подій, що відбуваються. За виявлення закономірностей у таких послідовностях можна з деякою часткою ймовірності прогнозувати появу подій у майбутньому, що дає змогу приймати правильніші рішення (Negnivitsky, 2002). Послідовністю називають впорядковану множину об'єктів. Для цього на множину має бути задано відношення порядку.

Тоді послідовність об'єктів можна описати в такому вигляді: $S = \{\ldots, i_p, \ldots, i_q\}$, де $p < q$. Наприклад, у випадку з аналізами такою послідовністю об'єктів може бути дата здачі аналізів. Така послідовність:

S = {(рівень гемоглобіну, 01.10.2017), (венозний тиск, 25.09.2017), (pH, 28.09.2017)}

Можна інтерпретувати як послідовність здачі аналізів однією людиною в різний час (спочатку поміряли венозний тиск, потім виміряли рівень pH, і вкінці рівень гемоглобіну).

Розрізняють два види послідовностей: з циклами і без циклів. У першому випадку допускається входження у послідовність одного і того самого об'єкта на різних позиціях:

$S = \{\ldots, i_p, \ldots, i_q, \ldots\}$, де $p < q$, $i_q = i_p$.

Кажуть, що транзакція T містить послідовність S, якщо $S \subseteq T$ і об'єкти, що входять у S, входять і в множину T зі збереженням відношення порядку. При цьому допускається, що в множині T між об'єктами з послідовності S можуть перебувати інші об'єкти.

Підтримкою послідовності S називають відношення кількості транзакцій, в яку входить послідовність S, до загальної кількості транзакцій. Послідовність є частою, якщо її підтримка перевищує мінімальну підтримку, задану користувачем: $Supp(S) > Supp_{min}$.

Завданням секвенційного аналізу є пошук всіх частих послідовностей: $L = \{S | Supp(S) > Supp_{min}\}$.

Основною відмінністю завдання секвенційного аналізу від пошуку асоціативних правил є встановлення відношення порядку між об'єктами множини I. Це відношення може бути визначено різними способами. Під час аналізу послідовності подій, що відбуваються в часі, об'єктами множини I є події, а відношення порядку відповідає хронології їх появи (Zhang, 2001).

Наприклад, під час аналізу послідовностей здачі аналізів у лікарні наборами є пакети аналізів, які здає пацієнт в різний час, а відношення порядку – це час здійснення цих аналізів:

D = {{(температура, артеріальний тиск, капілярний тиск), (pH, температура, пульс)}, {(рівень гемоглобіну в крові, температура), (артеріальний тиск, температура), (температура, венозний тиск)}, {(рівень гемоглобіну в крові)}}.

Звичайно, так виникає проблема ідентифікації пацієнтів. На практиці це вирішують введенням медичних карток, що мають унікальний ідентифікатор (табл. 3).

Табл. 3. Введення унікального ідентифікатора для множини аналізів

ID пацієнта	Послідовність здачі аналізів
0	(температура, артеріальний тиск, капілярний тиск), (pH, температура, пульс)
1	(рівень гемоглобіну в крові, температура), (артеріальний тиск, температура), (температура, венозний тиск)
2	(рівень гемоглобіну в крові)

Інтерпретувати таку послідовність можна так: пацієнт з ідентифікатором 0 спершу здав температуру, артеріальний та капілярний тиски, а з наступним своїм візитом здав рівень pH, показники температури та пульсу. Підтримка, наприклад, послідовності {(артеріальний тиск, температура)} становить 2/3, оскільки вона трапляється у пацієнтів з ідентифікаторами 0 та 1.

У багатьох прикладних областях об'єкти множини I природним чином поєднуються в групи, які своєю чергою також можуть об'єднуватися у більш загальні групи, і т. ін. Отже, виходить ієрархічна структура об'єктів.

Для прикладу такої ієрархії може бути така категоризація аналізів:

- тиск: артеріальний; венозний; капілярний.
- фізичні показники: температура; аналіз крові: рівень гемоглобіну; pH.

Наявність ієрархії змінює уявлення про те, коли об'єкт i присутній у транзакції T. Очевидно, що підтримка не окремого об'єкта, а групи, в яку він входить, більша

$$Supp(I_q) \geq Supp\ (i_j), \text{ де } i_j \in I_q.$$

Це пов'язано з тим, що під час аналізу груп підраховують не тільки транзакції, в які входить окремий об'єкт, але і транзакції, що містять всі об'єкти аналізованої групи. Наприклад, якщо підтримка *Supp* {артеріальний тиск, температура} = 2/3, то підтримка *Supp* {тиск, фізичні показники} = 2/3, оскільки об'єкти груп тиск і фізичні показники входять у транзакції з ідентифікаторами 0 і 1.

Використання ієрархії дає змогу визначити зв'язок, що входить у вищі рівні ієрархії, оскільки підтримка набору може збільшуватися, якщо підраховується входження групи, а не її об'єкта. Крім пошуку наборів, що часто трапляються у транзакціях, які своєю чергою складаються з об'єктів $F = \{i | i \in I\}$ або груп одного рівня ієрархії

$$F = \{I^g | I^g \in I^{g+1}\}.$$

Можна розглядати також змішані набори об'єктів і груп

$$F = \{i, I^g | i \in I^g \in I^{g+1}\}.$$

Це дає змогу розширити аналіз та отримати додаткові знання.

За ієрархічної структури об'єктів можна змінювати характер пошуку, змінюючи аналізований рівень. Очевидно, що чим більше об'єктів у множині I, тим більше об'єктів у транзакціях T і частих наборах. Це, своєю чергою, збільшує час пошуку й ускладнює аналіз результатів. Зменшити або збільшити кількість даних можна за допомогою ієрархічного уявлення аналізованих об'єктів. Переміщаючись вгору по ієрархії, узагальнюємо дані і зменшуємо їх кількість, і навпаки.

Недоліком узагальнення об'єктів є менша корисність отриманих знань, оскільки в цьому разі вони належать до груп, що не завжди несуть корисну інформацію. Для досягнення компромісу між аналізом груп і аналізом окремих об'єктів часто роблять так: спочатку аналізують групи, а потім, залежно від отриманих результатів, досліджують об'єкти, що зацікавили аналітика груп (Brin & Page, 2012). У будь-якому разі можна стверджувати, що наявність ієрархії в об'єктах і її використання в задачі пошуку асоціативних правил дає змогу виконувати більш гнучкий аналіз і отримувати додаткові знання.

У розглянутій задачі пошуку асоціативних правил наявність об'єкта в транзакції визначалося тільки його присутністю в ній ($i_j \in T$) або відсутністю ($i_j \notin T$). Часто об'єкти мають додаткові атрибути, як правило, чисельні. Наприклад, аналізи у транзакції мають атрибути: значення і тривалість. При цьому наявність об'єкта в наборі може визначатися не просто фактом його присутності, а і виконанням умови стосовно певного атрибуту. Наприклад, під час аналізу транзакцій, здійснених пацієнтами, цікавить не лише значення аналізу, а й наскільки цей показник є стабільним (довготривалим).

Для розширення можливостей аналізу за допомогою пошуку асоціативних правил у досліджувані набори можна додавати додаткові об'єкти. Загалом вони можуть мати природу, відмінну від основних об'єктів. Наприклад, у разі здачі аналізів можна ввести поле частоти здачі або симптоми, які передують для здачі саме цих аналізів.

Рішення завдання пошуку асоціативних правил, як і будь-якого завдання, зводиться до оброблення вихідних даних та отримання результатів. Оброблення вихідних даних виконують за певним алгоритмом Data Mining.

Результати, отримані при вирішенні цього завдання, прийнято представляти у вигляді асоціативних правил. У зв'язку з цим при їх пошуку виділяють два основних етапи: знаходження всіх великих наборів об'єктів; генерація асоціативних правил із знайдених великих наборів об'єктів.

Асоціативні правила мають такий вигляд:

Якщо (*умова*) *то* (*результат*),

де умова – зазвичай не логічний вираз (як у класифікаційних правилах), а набір об'єктів з множині I, з якими пов'язані (асоційовані) об'єкти, що входять у результат цього правила.

Наприклад, асоціативне правило:

Якщо (артеріальний тиск, pH), то (рівень гемоглобіну)

означає, що якщо пацієнт здає артеріальний тиск та рівень pH, то він здає і рівень гемоглобіну.

Як уже зазначено, в асоціативних правилах умова і результат є об'єктами множини I: *Якщо X то Y,*

де $X \in I$, $Y \in I$, $X \cup Y = \varphi$.

Основною перевагою асоціативних правил є їх легке сприйняття людиною і проста інтерпретація мовами програмування. Однак вони не завжди корисні (Sutton & Barto, 2008). Виділяють три види правил:

- *корисні правила* – містять дійсну інформацію, яка раніше була невідома, але має логічне пояснення. Такі правила можуть бути використані для прийняття рішень, що приносять вигоду;
- *тривіальні правила* – містять дійсну та легко зрозумілу інформацію, яка вже відома. Такі правила, хоча і можна пояснити, але не можуть принести будь-якої користі, оскільки відображають або відомі закони в досліджуваній області, або результати минулої діяльності. Іноді такі правила можуть використовуватися для перевірки виконання рішень, прийнятих на підставі попереднього аналізу;
- *незрозумілі правила* – містять інформацію, яка не може бути пояснена. Такі правила можуть бути отримані або на основі аномальних значень, або глибоко прихованих знань. Безпосередньо, такі правила не можна використовувати для прийняття рішень, позаяк їх нез'ясовність може призвести до непередбачуваних результатів. Для кращого розуміння потрібен додатковий аналіз.

Асоціативні правила будуються на основі великих наборів. Так, правила, побудовані на підставі набору F, є усіма можливими комбінаціями об'єктів, що входять у нього. Наприклад, для набору {артеріальний тиск, температура, пульс} можуть бути побудовані такі асоціативні правила:

- якщо (артеріальний тиск) то (температура);
- якщо (артеріальний тиск) то (пульс);
- якщо (артеріальний тиск) то (температура);
- якщо (артеріальний тиск) то (температура, пульс);
- якщо (температура, пульс) то (артеріальний тиск);
- і так далі.

Отже, кількість асоціативних правил може бути дуже великою і поганою для сприйняття людиною. До того ж, не всі з побудованих правил несуть в собі корисну інформацію. Для оцінки їх корисності вводять такі величини:

- *підтримка* (support) – показує, який відсоток транзакцій підтримує це правило;
- *достовірність* (confidence) – показує ймовірність того, що з наявності в транзакції набору X випливає наявність у ній набору Y;
- *покращення* (improvement) – показує, чи корисне це правило для дослідження.

Дані оцінки використовуються при генерації правил. Аналітик під час пошуку асоціативних правил задає мінімальні значення перерахованих величин (Roubus et al., 2001). Унаслідок цього ті правила, які не задовольняють ці умови, відкидаються і не включаються до рішення задачі.

Якщо об'єкти мають додаткові атрибути, які впливають на склад об'єктів у транзакціях, а отже, і в наборах, то вони повинні враховуватися в правилах, що генеруються. У цьому разі умовна частина правил буде містити не тільки перевірку наявності об'єкта у транзакції, але і складніші операції порівняння: більше, менше, включає та ін. Результатна частина правил також може містити твердження щодо значень атрибутів. Наприклад, якщо у показника розглядається актуальність, то правила можуть мати такий вигляд:

Якщо pH.актуальність > 10 днів то рівень гемоглобіну в крові.актуальність < 3днів.

Це правило свідчить про те, що пацієнт робив аналіз pH більше ніж 10 днів тому, то, ймовірно, його аналіз гемоглобіну в крові дійсний не більше ніж 3 дні.

Висновки. Завданням пошуку асоціативних правил є визначення наборів об'єктів, що часто трапляються, у великій множині об'єктів. Завданням секвенційного аналізу у пошук частих послідовностей. Основною відмінністю завдання секвенційного аналізу від пошуку асоціативних правил є встановлення відносини порядку між об'єктами. Наявність ієрархії в об'єктах і її використання в задачі пошуку асоціативних правил дає змогу виконувати більш гнучкий аналіз й отримувати додаткові знання. Результати рішення задачі представляються у вигляді асоціативних правил, умовна і заключна частина яких містить набори об'єктів.

Перелік використаних джерел

Brin, S., & Page, L. (2012). The anatomy of a large-scale hypertextual Web search engine: In *Seventh International World Wide Web Conference,* (pp. 23–29). Brisbane, Australia.

Johannes, R., Setnes, M., & Abonyic, J. (2001). *Learning Fuzzy Classification Rules from Labeled Data*. Delft: Information science. 320 p.

Negnivitsky, M. (2002). *Artificial Intelligence – A Guide to Intelligent Systems*. Addison-Wesley: Pearson Education Limited. 230 p.

Sutton, R. S., & Barto, A. G. (2008). *Reinforcement Learning. An Introduction*. London: MIT Press, Cambridge. 268 p.

Zhang, L. (2001). *Comparation of Fuzzy c-means Algorithm and New Fuzzy Clustreing and Fuzzy Merging Algorithm*. Nevada: Computer Science Departament, University of Nevada. 328 p.

И. И. Желизняк

Национальный университет "Львовская политехника", г. Львов, Украина

ПРАВИЛА ПОСТРОЕНИЯ АССОЦИАТИВНЫХ ПРАВИЛ НА ПРИМЕРЕ ФИЗИЧЕСКИХ ПОКАЗАТЕЛЕЙ ПАЦИЕНТА

Охарактеризованы правила построения ассоциативных правил. Обозначены объекты, составляющие исследуемый набор. Построены ассоциативные правила для назначения анализов пациенту. Приведены наиболее распространенные термины интеллектуального анализа данных. Рассмотрено множество транзакций, которые доступны для медицинского анализа пациента. Описана последовательность объектов и заданное отношение порядка. Указано минимальное значение поддержки множества и отсеяны ассоциативные правила на основе этого значения. Охарактеризованы величины полезности ассоциативных правил, с помощью которых устанавливается важность того или иного ассоциативного правила. Выявлено, что правильная оценка полезности ассоциативного правила влияет на объем и быстродействие доступа к информации. Введен уникальный идентификатор для исследуемого набора анализов пациента. Отмечены дополнительные многочисленные атрибуты исследуемых объектов. Охарактеризованы транзакции, которые содержат дополнительные атрибуты и операции не только наличия, а также сравнения. Установлено различие между ассоциативными правилами и секвенциальным анализом. Оценено влияние построения ассоциативных правил при исследовании предметной области и интеллектуальном анализе данных. Полученные результаты будут использованы в дальнейших исследованиях в данной предметной области.

Ключевые слова: секвенциальный анализ; data mining; медицина; биоинформатика; выявление закономерностей.

I. J. Zheliznyak

Lviv Polytechnic National University, Lviv, Ukraine

SOME REGULATIONS FOR CONSTRUCTING ASSOCIATIVE RULES ON THE EXAMPLE OF PATIENT'S PHYSICAL CHARACTERISTICS

The authors have investigated one of the important sections of the data mining process. The concept of bioinformatics is considered. The objects included in the study set are defined. The object of the study is the patient's physical parameter and the value of this indicator. The set of transactions, the information about which is available for analysis, is defined. It describes the value of the support for the studied sets, as well as the minimum support is set. We have described and characterized sequences of investigated objects. Various types of sequences are described, including cycles and without cycles. The support for the sequence of investigated objects as the ratio of the number of transactions, which includes the studied sequences, to the total number of transactions is characterized. Examples of sequence analysis and ordering in this subject area are given. A unique identifier has been entered for each patient. The hierarchical structure is described on the example of the patient's physical characteristics. The advantages and disadvantages of using the hierarchical structure are presented. We have proved that the use of this method allows more flexible analysis and additional knowledge. Additional numerical attributes of the investigated objects are described. The main stages of the formation and presentation of associative rules are described. We have defined such types of associative rules as useful, trivial and obscure ones. The associative rules for the studied set of patient's physical indicators are derived. The values of the utility of associative rules, such as support, reliability, improvement, etc. are indicated. We have also characterized the rules for selection and filtering associative rules. Moreover, the authors have introduced more complex comparisons in the conditional part of the associative rule. We used regulations for constructing associative rules and sequencing analysis when conducting the study. These results will be used in further research to identify the patterns of symptoms and diagnoses of the patient.

Keywords: sequential analysis; data mining; medicine; bioinformatics; detection of regularities.