

BIG DATA: ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ РИНКОВОГО КОШИКУ

У тезах представлено результати застосування методів інтелектуального аналізу даних, зокрема алгоритму *Apriori*, для аналізу ринкового кошика на основі реального набору транзакцій супермаркету, що дозволило виявити значущі асоціативні правила та патерни спільних покупок, які можуть бути використані для оптимізації бізнес-стратегій у різних сферах.

Ключові слова: *Big Data, Market Basket Analysis, асоціативні правила, алгоритм Apriori, dataset Supermarket.*

Інтелектуальний аналіз даних (Data mining) – це процес, у якому алгоритми використовуються для аналізу даних і пошуку цікавих і нетривіальних шаблонів (patterns), які можуть бути корисними, скажімо для бізнесу [1]. Методи пошуку асоціативних правил не потребують вибору атрибуту класу, на відміну від задач регресії, чи класифікації, – усі атрибути набору даних вважаються атрибутами ознак. Методи побудови асоціативних правил призначені для пошуку комбінацій значень атрибутів, на базі яких за комбінацією значень атрибутів первинної множини (так званих «умов») можна спрогнозувати значення атрибутів вторинної множини (так званих «наслідків»). Одним із найпопулярніших методів, які використовуються для цієї мети, є аналіз ринкового кошика, також відомий як аналіз спорідненості (табл. 1). Цей метод передбачає аналіз записів транзакцій (припустимо закупок у супермаркеті), щоб виявити цікаві моделі купівлі, які можна використовувати для покращення бізнес-стратегій, таких як макетування пропозиції продуктів.

Таблиця 1.

Основні галузі застосування «аналізу ринкового кошику»

Галузь	Застосування «аналізу ринкового кошику»
Роздрібна торгівля	Визначення комбінацій продуктів, які часто купують, і створення рекламних акцій, або стратегій перехресних продаж.
Електронна комерція	Пропозиція клієнтам додаткових продуктів та покращення взаємодії з ними
Ресторанний бізнес	Визначте, які пункти меню часто замовляють разом, і створіть пакети (комплекси) страв або рекомендації щодо меню.
Охорона здоров'я	Розуміння, які ліки (процедури) часто призначають разом, і визначення моделі поведінки пацієнтів, або результати лікування
Банківська справа/фінанси	Визначте, які продукти чи послуги клієнти часто використовують разом, і створіть цільові маркетингові кампанії або пакетні угоди
Телекомунікації	Зрозумійте, які продукти чи послуги часто купуються разом, і створіть комплексні пакети послуг, які збільшать дохід і покращать взаємодію з клієнтами

Знаходження (видобуток) правил асоціації зазвичай розглядається як двоетапний процес:

1) по-перше, ідентифікуються всі часті набори елементів (атрибутів). За визначенням, ці набори елементів будуть зустрічатися принаймні так само часто, як попередньо встановлена мінімальна кількість підтримки, відома як *min sup*;

2) другий крок передбачає генерацію правил асоціації з цих частих наборів елементів. Ці правила повинні задовольняти як мінімальну підтримку, так і мінімальну довіру (*confidence*).

Weka, один з найкращих сучасних програмних продуктів для інтелектуального аналізу даних, поставляється з низкою реальних наборів даних у каталозі «data» інсталяції Weka [2]. Це дуже зручно, оскільки дає можливості досліджувати та експериментувати з цими добре відомими проблемами та дізнаватися про різні методи, доступні в Weka. Розглянемо зокрема реальний набір даних Supermarket (data/supermarket.arff) [3]. Це набір даних про діючу торгову точку. Всі атрибути номінальні, і кожен екземпляр (інстанція) представляє транзакцію реального клієнта в супермаркеті, придбані продукти та

здіянні відділи. Це саме те, що називають Big Data: дані містять 4'627 екземплярів (інстанцій) і 217 атрибутів для кожної з них (рис. 1).

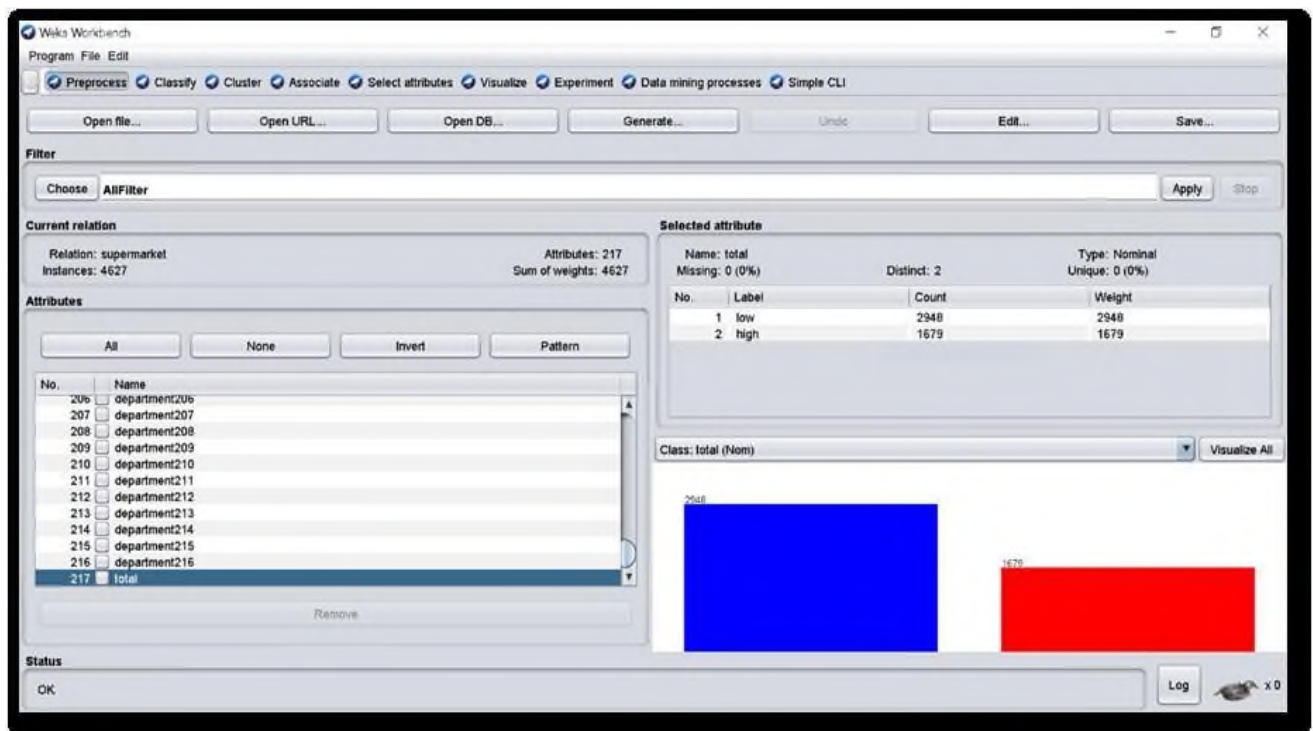


Рис. 1. Набір даних Supermarket у вікні-модулі Explorer Weka

Дані не нормалізовані. Кожен атрибут є двійковим і має значення («t» для істини), або не має значення (знак «?» для відсутності певного значення). Існує також номінальний атрибут класу під назвою «total», який вказує, чи була транзакція (закупівля) за ціною меншою за 100 доларів (low) чи більшою за 100 доларів (high). Тут нема зацікавлення у створенні прогностної моделі для номінального класу «total». Натомість нас цікавить, які товари були найчастіше придбані разом. Є зацікавленість у пошуку корисних закономірностей у цих даних, які можуть, або не можуть, бути пов'язані з прогнозованим атрибутом.

Обираємо вкладку «Associate» у Weka Workbench. Застосовуємо алгоритм «Апріорі». Це найвідоміший метод вивчення правил асоціації, тому що він, можливо, був першим (Агравал і Срікант у 1994 році) і загальне визнання як дуже ефективний. Середній комп'ютер витрачає на пошук десяти асоціативних правил в досліджуваному наборі приблизно 3 секунди (дивись звіт під кнопкою Log). Як результат пошуків правил маємо таке вікно (рис. 2).

У вікні «Association output» можна побачити, що алгоритм представив десять правил, отриманих із набору даних супермаркету. Алгоритм налаштований на зупинку на десяти правилах за замовчуванням; втім можна клацнути на назві алгоритму та налаштувати його для пошуку та звітування більшої чи меншої кількості правил, змінивши значення «numRules».

Виявлені правила були такі:

- 1) biscuits=t frozen foods=t fruit=t total=high 788 ==> bread and cake=t 723 conf:(0.92)
- 2) baking needs=t biscuits=t fruit=t total=high 760 ==> bread and cake=t 696 conf:(0.92)
- 3) baking needs=t frozen foods=t fruit=t total=high 770 ==> bread and cake=t 705 conf:(0.92)
- 4) biscuits=t fruit=t vegetables=t total=high 815 ==> bread and cake=t 746 conf:(0.92)
- 5) party snack foods=t fruit=t total=high 854 ==> bread and cake=t 779 conf:(0.91)
- 6) biscuits=t frozen foods=t vegetables=t total=high 797 ==> bread and cake=t 725 conf:(0.91)
- 7) baking needs=t biscuits=t vegetables=t total=high 772 ==> bread and cake=t 701 conf:(0.91)
- 8) biscuits=t fruit=t total=high 954 ==> bread and cake=t 866 conf:(0.91)
- 9) frozen foods=t fruit=t vegetables=t total=high 834 ==> bread and cake=t 757 conf:(0.91)
- 10) frozen foods=t fruit=t total=high 969 ==> bread and cake=t 877 conf:(0.91)

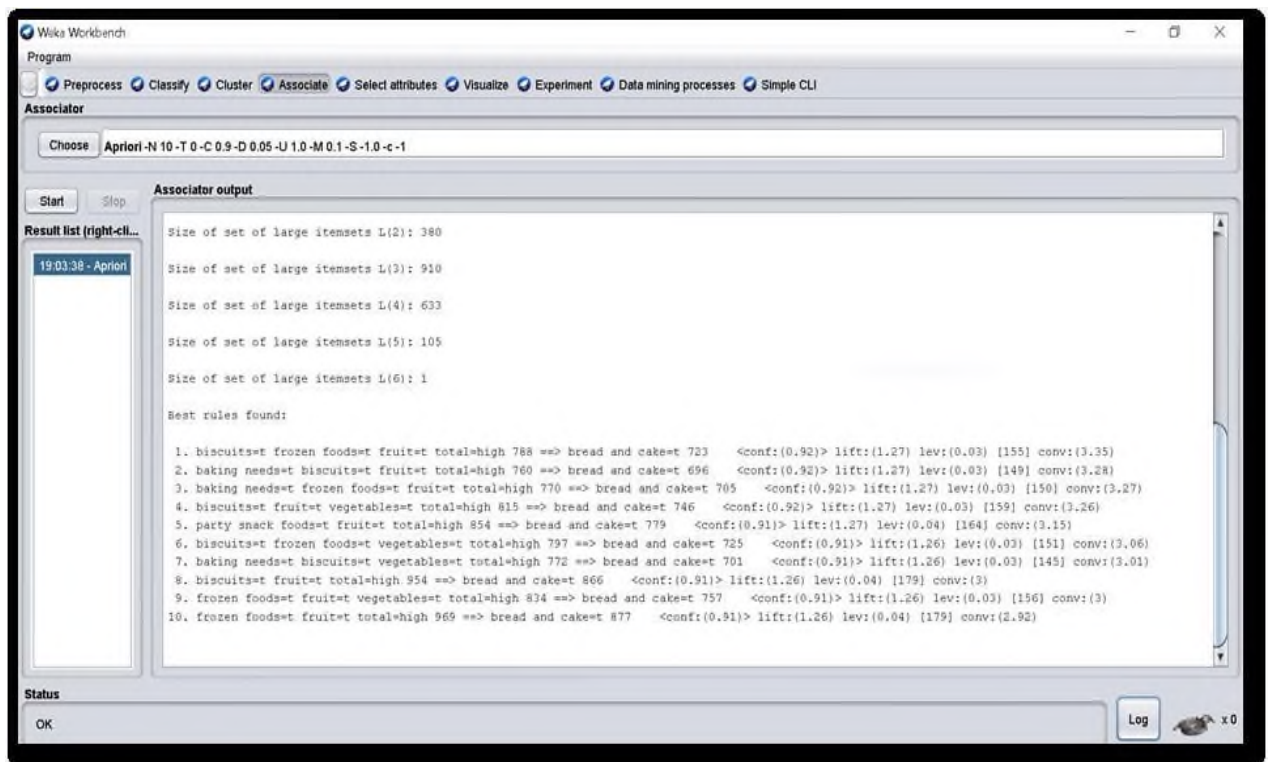


Рис. 2. Десять асоціативних правил знайдених алгоритмом Apriori у наборі даних Supermarket

Правила представлені у форматі antecedent => consequent (умова => наслідок). Число, пов'язане з умовою, є абсолютним охопленням (support) у наборі даних (у цьому випадку це число з можливих 4'627). Число поруч із наслідком – це абсолютна кількість випадків, які збігаються з умовою і наслідком. Число в дужках наприкінці підтверджує правило (кількість умов поділена на кількість відповідних наслідків). Ви бачите, що для вибору правил було використано обмеження в 91 %, згадане у вікні «minMetrics», отже жодне правило не має довірчого інтервалу (confidence) менше ніж 0,91.

Ось лише кілька спостережень, які самі «впадають в око»:

- бачимо, що всі представлені правила мають наслідок «хліб та випічка»;
- усі представлені правила вказують на високу загальну суму транзакції (стосуються класу «high»);
- «печиво» та «заморожені продукти» фігурують у багатьох представлених правилах.

Аналітики повинні бути дуже обережними щодо тлумачення правил асоціації. Це асоціації (що є синонімом кореляцій), які необов'язково є причинно-наслідковими зв'язками. Крім того, короткий антецедент (умова) є більш міцним, ніж довгий антецедент, який, швидше за все, буде крихким. Наприклад, якщо вас цікавить загальна сума, витрачена покупцем, можна намагатися переконувати людей, які купують печиво, заморожені продукти та фрукти, купувати також і хліб і тістечка, щоб вони призвели до високої загальної суми транзакції (правило № 1 вище). Така думка може здатися корисною лише на перший погляд, але вона є помилковим міркуванням. Комбінація продуктів не призводить до високого підсумку витрат клієнта, вона сама є наслідком високих витрат. Іншими словами, асоціативні правила ніколи не діють у зворотному напрямку.

References

1. Agarwal S. Data Mining: Data Mining Concepts and Techniques. *2013 International Conference on Machine Intelligence and Research Advancement (ICMIRA)*, Katra, India, 21–23 December 2013. 2013. DOI: 10.1109/icmira.2013.45.
2. Duran Z., Akargöl İ., Doğan T. Data Mining, Weka Decision Trees. *Orclever Proceedings of Research and Development*. 2023. Vol. 3, no. 1. P. 401–416. DOI: 10.56038/oprd.v3i1.376.
3. Sharma A., Babbar H. Analysis of Data Mining Algorithms in Market Basket Analysis. *2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT)*, Gharuan, India, 5–6 May 2023. 2023. DOI: 10.1109/incacct57535.2023.10141816.

BIG DATA: INTELLIGENT MARKET BASKET ANALYSIS

The theses present the results of applying data mining methods, in particular the Apriori algorithm, for market basket analysis based on a real set of supermarket transactions. This approach made it possible to identify significant association rules and patterns of joint purchases that can be used to optimize business strategies in various fields.

Key words: *Big Data, Market Basket Analysis, association rules, Apriori algorithm, Supermarket dataset.*

Відомості про автора / Information about the Author

Геннадій Чуйко, д-р фіз.-мат. наук, професор, професор кафедри комп'ютерної інженерії, Чорноморського національного університету імені Петра Могили, м. Миколаїв, Україна. E-mail: genchuiko@gmail.com, orcid: <https://orcid.org/0000-0001-5590-9404>.

Gennady Chuiko, Doctor of Physical and Mathematical Sciences, Professor, Professor of the Department of Computer Engineering, Petro Mohyla Black Sea National University, Mykolaiv, Ukraine. E-mail: genchuiko@gmail.com, orcid: <https://orcid.org/0000-0001-5590-9404>.

© Г. Чуйко
20.05.2025.

Стаття отримана редакцією

