Інформаційні технології

УДК 004.891

Тарасюк Тарас Сергійович
студент
Навчально-наукового комплексу «Інститут прикладного системного аналізу»
Національного технічного університету України
"Київський політехнічний інститут імені Ігоря Сікорського"
Тарасюк Тарас Сергеевич
студент
Учебно-научного комплекса «Институт прикладного системного анализа»
Национального технического университета Украины
"Киевский политехнический институт имени Игоря Сикорского"
Tarasiuk Taras
Student of the
ESC «Institute for Applied Systems Analysis» of the
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute"

МЕТОДИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ В РОЗДРІБНІЙ ТОРГІВЛІ
МЕТОДЫ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ В РОЗНИЧНОЙ ТОРГОВЛЕ
DATA MINING METHODS IN RETAIL

Анотація. *Розглянуто основні підходи та методи асоціативного та регресійного аналізу, розроблено та реалізовано у вигляді програмного модуля архітектуру СППР для аналізу транзакцій роздрібного підприємства, моделювання та прогнозування прибутку та попиту.*

Ключові слова: *інтелектуальний аналіз даних, роздрібна торгівля, асоціативний аналіз, регресійний аналіз, система підтримки прийняття рішень.*

Аннотация. *Рассмотрены основные подходы и методы ассоциативного и регрессионного анализа, разработана и реализована в виде программного модуля архитектура СППР для анализа транзакций розничного предприятия, моделирования и прогнозирования прибыли и спроса.*

Ключевые слова: *интеллектуальный анализ данных, розничная торговля, ассоциативный анализ, регрессионный анализ, система поддержки принятия решений.*

Summary. *Approaches and methods of associative and regression analysis are surveyed, DSS architecture was developed and implemented in the form of the software product for the analysis of retail transactions, modeling and forecasting of profits and demand.*

Key words: *data mining, retail, associative analysis, regression analysis, decision support system.*

Вступ. За останні десятиліття підприємства досягли значних успіхів у сфері роздрібної торгівлі. Це стало можливим завдяки оптимізації збору, збереження та обробки інформації про клієнтів та їх покупки. На даний момент існує безліч систем, що дозволяють автоматизувати ці процеси та провести первинний аналіз отриманих даних. Тому виникають завдання, пов'язані з необхідністю обробки великих масивів даних з метою пошуку нових закономірностей, встановлення і виявлення нових знань, що можуть бути вирішені у вигляді нових аналітичних систем, які базуючись на методах інтелектуального аналізу даних, зможуть надати користувачу обґрунтовану

інформацію для прийняття рішень про асортимент, розміщення та комплектацію товарів і стратегію розвитку підприємства.

Мета дослідження. Розробка структури СППР та її реалізація у вигляді програмного продукту для побудови асоціацій, моделювання та прогнозування основних бізнес-процесів у роздрібній торгівлі.

Об'єкт дослідження. Статистичні ряди даних, що описують бізнес-процеси роздрібної торгівлі та потребують аналітичної обробки та виділення знань про їхні взаємозв'язки, необхідні в процесі прийняття рішень.

Предмет дослідження. Моделі та методи інтелектуального аналізу даних: алгоритм Apriori для генерації асоціативних правил, моделі регресійного аналізу з прогнозування часових рядів: AR, ARMA, ARIMA та моделей у вигляді тренду.

Методи дослідження. Спираються на теорію асоціативного аналізу та теорію моделювання часових рядів.

Постановка задачі:

1. Провести огляд методів та підходів асоціативного та регресійного аналізу до задач пошуку взаємозалежностей, моделювання і прогнозування.
2. Спроектувати архітектуру СППР, призначеної для аналізу, знаходження залежностей, моделювання та прогнозування основних бізнес-процесів процесів підприємства роздрібної торгівлі.
3. Реалізувати алгоритм асоціативного аналізу Apriori та методи регресійного аналізу: AR, ARMA, ARIMA, моделі у вигляді поліноміального тренду у формі програмного продукту.
4. Застосувати розроблену СППР до реальних даних.
5. Провести аналіз отриманих результатів.

Алгоритм Apriori

Алгоритм Apriori використовує одну із властивостей підтримки: підтримка будь-якого набору об'єктів не можу перевищувати підтримки будь-якого із його підмножин:

$$Supp_F \leq Supp_E, E \subset F.$$

Алгоритм Apriori визначає набори, що часто зустрічаються за декілька етапів. На i–му етапі визначаються всі часті i–елементні набори. Кожен етап складається із двох кроків: формування кандидатів і підрахунку підтримки кандидатів.

Позначення, що використовуються в алгоритмі:

$L_k -$ множина k–елементних частих наборів, чия підтримка не менша заданої користувачем. Кожен член множити має набір впорядкованих ($i_j < i_p$, якщо $j < p$) елементів F і значення підтримки набору $Supp_F > Supp_{min}$:

$$L_k = \{(F_1, Supp_1), (F_2, Supp_2), \dots, (F_q, Supp_q)\},$$

де $F_j = \{i_1, i_2, \dots, i_k\}$;

$C_k -$ множина кандидатів $k -$ елементних наборів потенційно частих. Кожен член множини має набір впорядкованих ($i_j < i_p$, якщо $j < p$) елементів F і значення підтримки набору $Supp$.

Алгоритм Apriori покроково:

Крок 1. Присвоїти $k = 1$ і виконати відбір усіх 1–елементних наборів, у котрих підтримка більша мінімально заданого користувачем значення $Supp_{min}$.

Крок 2. $k = k + 1$.

Крок 3. Якщо не вдається створити k–елементні набори, то завершити алгоритм, інакше перейти на наступний крок.

Крок 4. Створити множину k–елементних наборів кандидатів в часті набори. Для цього необхідно об'єднати в k–елементні кандидати $(k-1)$ – елементні часті набори. Кожен кандидат $c \in C_k$ буде формуватися шляхом додавання до $(k-1)$ – елементному частому набору – p елемента із іншого $(k-1)$ – елементного частого набору – q. При чому додається останній елемент набору q, який по порядку вище, чим останній елемент набору p. При цьому усі перші $k-2$ елементи обох наборів однакові.

Крок 5. Для кожної транзакції T із множини D вибрати кандидатів C_t із множини C_k, що присутні в транзакції T. Для кожного набору із побудованої множини C_k видалити набір, якщо хоча б одна із його $(k-1)$ підмножин не є частим, тобто відсутній в множині L_{k-1}.

Крок 6. Для кожного кандидата із множини C_k збільшити значення підтримки на одиницю.

Крок 7. Вибрати лише кандидатів L_k із множини C_k, у яких значення підтримки більше заданого користувачем $Supp_{min}$.

Результатом роботи алгоритму є об'єднання всіх множин L_k для всіх k

Моделі прогнозування

1. Авторегресійна модель.

   Авторегресійна модель порядку p має наступний вигляд:

$$y(k) = a_0 + a_1 y(k-1) + a_2 y(k-2) + \cdots + a_p y(k-p) + \varepsilon(k)$$

2. Модель авторегресії з ковзним середнім.

   Модель авторегресії з ковзним середнім має вигляд:

$$y(k) = a_0 + a_1 y(k-1) + a_2 y(k-2) + \cdots + a_p y(k-p) + \varepsilon(k) - b_1 \varepsilon(k-1) - b_2 \varepsilon(k-2) - \cdots - b_q \varepsilon(k-q).$$

де $p -$ порядок авторегресійної частини моделі, $q -$ порядок частини ковзного середнього.

3. Модель у вигляді поліноміального тренду.

Модель у вигляді поліноміального тренду можна записати з використанням детермінованих функцій від часу у вигляді:

$$y(k) = a_0 + d_1 k + d_2 k^2 + \cdots + d_m k^m + \varepsilon(k)$$

4. Модель авторегресії з інтегрованим ковзним середнім.

$$\tilde{y}(k) = d^m y(k),$$

де $dy(k) = y(k) - y(k-1)$.

$$\tilde{y}(k) = a_0 + a_1\tilde{y}(k-1) + a_2\tilde{y}(k-2) + \cdots + a_p\tilde{y}(k-p) + \varepsilon(k) - b_1\tilde{\varepsilon}(k-1) - b_2\tilde{\varepsilon}(k-2) - \cdots - b_q\tilde{\varepsilon}(k-q)$$

Архітектура СППР

Архітектура СППР, що зображена на рис. 1, складається з наступних блоків:

- блок завантаження даних;
- блок попереднього аналізу;
- блок генерації асоціативних правил;
- блок побудови моделей;
- блок прогнозування;
- блок обчислення оцінок;
- блок виведення результатів.

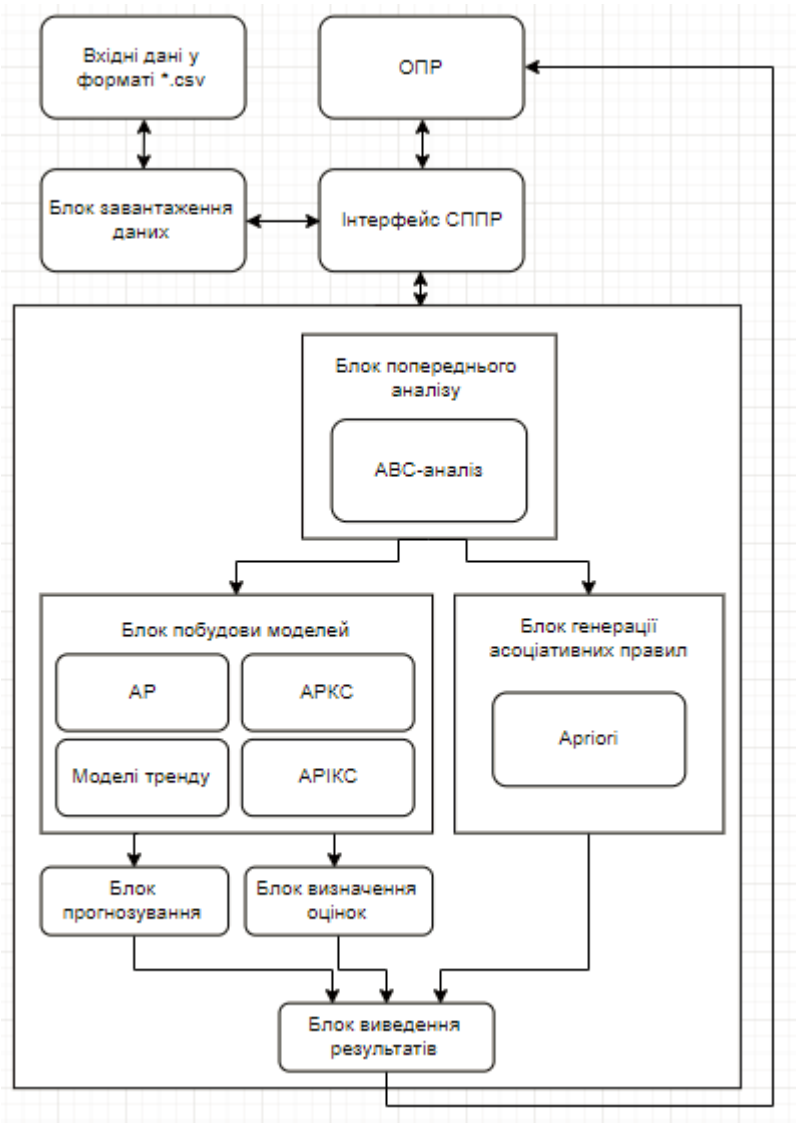

Рис. 1. Блок-схема СППР

Застосування СППР

Побудовану СППР застосовано для роздрібного магазину на основі даних про товари, категорії товарів та транзакції за період першого та другого кварталів 2016 року.

В результаті проведеного АВС-аналізу, категорії товарів поділені на три множини: A, B та C, що приносять 80%, 15% і 5% прибутку відповідно. Для множини категорій товарів, згенеровані асоціативні правила, що зображені у табл. 1. Відповідно до отриманих правил, видно, що покупці найчастіше

купують товари разом з категорій з кодом 141 та 142, також в більшості випадків з цими категоріями разом купують товари з категорій 151 та 143.

Таблиця 1

Асоціативні правила для категорій товарів із множини А

Правило	Підтримка	Впевненість	Поліпшення
{ 141 } → { 142 }	0,029	0,4284	4,8601
{ 141 } → { 143 }	0,0247	0,3657	4,364
{ 062 } → { 182 }	0,0218	0,3569	3,6346
{ 142 } → { 143 }	0,0312	0,3536	4,2193
{ 962 151 } → { 152 }	0,0133	0,4272	6,9658
{ 962 151 } → { 163 }	0,014	0,4114	4,6005
{ 962 152 } → { 151 }	0,0133	0,3678	6,7564
{ 152 143 141 } → { 142 }	0,0055	0,6778	7,6891
{ 143 141 062 } → { 142 }	0,0051	0,676	7,6686
{ 151 143 142 } → { 141 }	0,0058	0,5829	8,6172

Для побудови моделі попиту розглянемо дані щотижневого продажу однієї з категорій товарів - 141:

АР:

$$Y = 8149{,}44 + 0{,}1695 * y(k-2) + 0{,}11 * y(k-3) + 0{,}10 * y(k-4) + 0{,}43 * y(k-5)$$

АРКС:

$$Y = 8149{,}44 + 0{,}16 * y(k-2) + 0{,}11 * y(k-3) + 0{,}10 * y(k-4) + 0{,}43 * y(k-5) + + 1109{,}65 - 1{,}72 * \varepsilon(k-1) - 0{,}01 * \varepsilon(k-3)$$

Тренд першого порядку:

$$Y(k) = 36489{,}25 + 290{,}83 * k$$

АРІКС

$$Y = -0{,}3 - 0{,}3 * y(k-2) - 0{,}3 * \hat{y}(k-3) - 0{,}3 * \hat{y}(k-4) - 0{,}3 * \hat{y}(k-5) + +1963{,}20 - 0{,}25 * \varepsilon(k-1) + 0{,}29 * \varepsilon(k-3) + y(k-1)$$

Відповідно до наведених статистик у табл. 2, найкращою виявилась модель у вигляді тренду першого порядку.

Таблиця 2

Значення статистик для навчальної та тестової вибірок

	RMSE		**Коефіцієнт детермінації**	
Метод	**Навчальна**	**Тестова**	**Навчальна**	**Тестова**
AP	2844,9	4009,33	0,233	0,044
APKC	2631,4	4208,75	0,211	0,238
Тренд 1-го порядку	2596,74	3756,855	0,315	0,017
APIKC	3412,4	5414,98	0,21	0,238

Висновки. В статті розглянуто підходи до аналізу даних роздрібної торгівлі, пошуку асоціацій, моделювання та прогнозування попиту на категорії товарів.

Спроектовано архітектуру СППР для аналізу транзакцій, пошуку залежностей між товарами та їх категоріями, моделювання та прогнозування попиту. Реалізований метод Apriori для пошуку асоціативних правил та методи регресійного аналізу для побудови моделей і прогнозування: AR, ARMA, ARIMA, моделей у вигляді тренду.

Описані методи реалізовані в програмному модулі зі створеним користувацьким інтерфейсом.

Програмний продукт застосовано на прикладі даних роздрібного магазину. Згенеровано асоціативні правила для категорій товарів, виявилось, що покупці найчастіше купують разом товари з категорій 141, 142, 143 та 151. А при моделюванні і прогнозуванні попиту на категорію товару з кодом 141, найкращою виявилась модель у вигляді тренду 1-го порядку

$$Y(k) = 36489{,}25 + 290{,}83 * k$$

Література

1. Паклин Н.Б. Бизнес-аналитика: от данных к знаниям / Паклин Н.Б., Орешков В.И. – СПб.: Питер. - 2013. - 704 с.
2. Бідюк П. І. Часові ряди: моделювання і прогнозування / Бідюк П.І., Савенков О. І., Баклан І.В. – К.: ЕКМО, 2003. – 144 с.
3. Бокс Дж., Анализ временных рядов. Прогноз и управление / Дж. Бокс, Г. Дженкинс - М.: Мир, 1974. - 402 с.
4. Zhang C. Association rule mining: models and algorithms / C. Zhang, S. Zhang. – Berlin: Springer-Verlag. – 2002. – 238 p.
5. Shin Y.C. Intelligent systems: modeling, optimization, and control / C.Y. Shin, C. Xu. – Boca Raton: CRC Press, 2009. – 456 p.
6. Adamo J.-M. Data mining for association rules and sequential patterns: sequential and parallel algorithms / J.-M. Adamo. – New York: Springer-Verlag. – 2001. – 259 p.