# ПОРІВНЯЛЬНИЙ АНАЛІЗ ЗАСОБІВ DATA MINING У СКБД SQL SERVER ТА ORACLE

**УДК 004.652**

**ФІСУН Микола Тихонович**
доктор технічних наук, професор, завідувач кафедри інтелектуальних інформаційних систем, Чорноморський державний університет імені Петра Могили, м. Миколаїв, Україна.
**Наукові інтереси:** інтелектуальні інформаційні системи та CASE-засоби їх створення, системи автоматизованого проектування, бази даних та бази знань.
mykola.fisun@gmail.com

**ДАВИДЕНКО Євген Олександрович**
кандидат технічних наук, в.о. доцента (б.в.з.) кафедри інтелектуальних інформаційних систем, Чорноморський державний університет імені Петра Могили, м. Миколаїв, Україна.
**Наукові інтереси:** програмні продукти та технології Microsoft, бази даних та бази знань, системи автоматизованого проектування, Web-технології, системи підтримки прийняття рішень, системний аналіз.
genik.davydenko@gmail.com

**КРАЙНИК Олексій Михайлович**
магістр, Чорноморський державний університет імені Петра Могили, м. Миколаїв, Україна.
**Наукові інтереси:** Data Mining, OLAP-технології, інтелектуальний аналіз даних.
alex.krainyk@gmail.com

**ВСТУП**

Елементи автоматичної обробки і аналізу даних, що називають Data Mining (добування знань) стають невід'ємною частиною концепції електронних сховищ даних та організації інтелектуальних обчислень. Як правило, Data Mining являє собою обчислювальний процес виявлення патернів у великих наборах даних з використанням методів штучного інтелекту, машинного навчання, статистики та систем керування базами даних (СКБД). Існує велика кількість програмних продуктів, таких як, IBM DB2 Intelligent Miner, Microsoft Analysis Services, Oracle DM, RapidMiner, які дозволяють використовувати технологію Data Mining з корпоративними базами даних (тобто такими базами даних, що об'єднують в тому чи іншому вигляді усі необхідні данні та знання про автоматизовану систему). Особливо слід виділити дві СКБД, які користуються найбільшою популярністю серед клієнтів та розробників – Oracle Database і MS SQL Server.На сьогоднішній день СКБД Microsoft SQL Server [1] та Oracle є основними програмними продуктами, які надають можливість виконувати обробку даних за допомогою засобів Data Mining [1].

Засоби Data Mining для кожної з вказаних СКБД представляють окремий набір інструментів. Для SQL Server це Analysis Services, а також допоміжні служби. Для Oracle Database такі засоби представлені Oracle Data Mining, Oracle SQL Developer та Oracle Data Miner.

**Аналіз останніх досліджень**. Вирішення задач, що потребують аналізу існуючих даних для прийняття рішень, засобами Data Mining отримують широке розповсюдження. Прикладами сфер використання інтелектуального аналізу даних є вирішення проблем промисловості [2], торгівлі [3], освітнього процесу [4], інформаційних ресурсів [5] та ін.

У вказані вище джерелах наводиться або власна реалізація аналізу даних, або, при їх обґрунтуванні, відбувається посилання лише на результати, отримані за допомогою тільки одного програмного інструменту. Тим не менш, в залежності від вибору програмного засобу, остаточні підсумки можуть відрізнятися та значно впливати на подальші дії, а, значить, і на кінцевий результат, наприклад, отриманий прибуток, кількість клієнтів тощо. Саме тому актуальною є проблема порівняння результатів роботи різних реалізацій засобів Data Mining при їх роботі з однаковим набором даних.

**Мета роботи**. Метою роботи є дослідження функціональних особливостей технології Data Mining, що інтегрована в СКБД SQL Server і Oracle та порівняльний аналіз результатів використання аналогічних методів класифікації, кластеризації та асоціації.

## ОСНОВНИЙ МАТЕРІАЛ

Для порівняння використаємо однакові набори даних, які містяться в навчальній базі від компанії Microsoft – AdventureWorks, тестові дані Oracle та статистичні бази даних, що знаходяться у відкритому доступі.

Класифікація, кластеризація та асоціація є найбільш використовуваними задачами Data Mining, що дозволяють розглядати досить великий обсяг інформації та різко скорочувати, стискати об'ємні масиви даних, робити їх компактними і наочними.

В SQL Server та Oracle Database є можливість порівнювати отримані моделі між собою та обирати найбільш точну. Для реалізації поставленої мети обираємо декілька алгоритмів класифікації, що використовують однаковий набір даних – дерево рішень, SVM, GLM, класифікатор Баєса.

На рис. 1 приведено графік точності використаних моделей (за допомогою інструментів SQL Server), в порівнянні з ідеальною моделлю. Подібний функціонал також реалізований в Oracle (рис. 2).

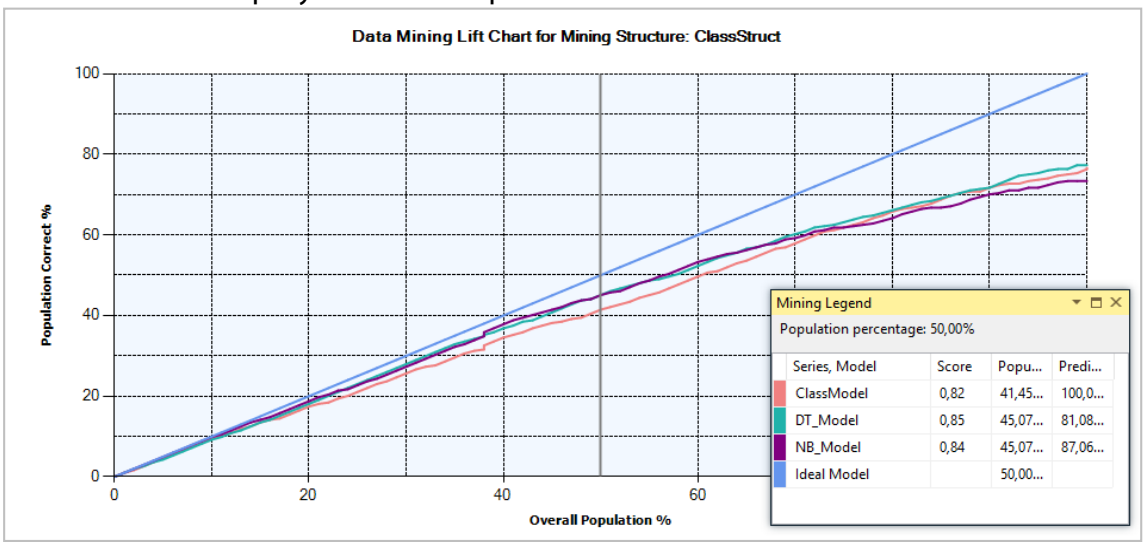


*Рис. 1. Графік точності моделей в SQL Server*

Діаграма точності прогнозів графічно відображає, яке поліпшення дасть модель інтелектуального аналізу даних в порівнянні з випадковим припущенням, а також вимірює зміну в термінах оцінки точності. Порівняння цих оцінок для різних фрагментів набору даних і різних моделей дозволяє вибрати кращу модель, а також визначити відсоток випадків в наборі даних, в яких прогнози моделі виявляться корисними.

Дана діаграма дозволяє порівняти точність прогнозів для декількох моделей, що мають однаковий прогнозований атрибут. Крім того, можна отримувати точність прогнозування як для одиничного випадку (середнє арифметичне значення прогнозованого атрибута), так і для всіх випадків (всі значення зазначеного атрибута).

Якщо не вказано стан прогнозованого стовпця, то можна створити діаграму наступного типу, показаного на попередній діаграмі (рис. 1). На цій діаграмі показана точність моделі для всіх станів прогнозованого атрибута. Наприклад, вона покаже наскільки точно модель передбачає клієнтів, які збираються страхуватись.

Вісь *X* така ж, як і на діаграмі з заданим прогнозованим стовпчиком, але вісь *Y* тепер представляє відсоток прогнозів, які є правильними. Тому ідеальна лінія проходить по діагоналі і показує, що при 50% даних модель вірно прогнозує 50% варіантів, тобто очікуваний максимум.

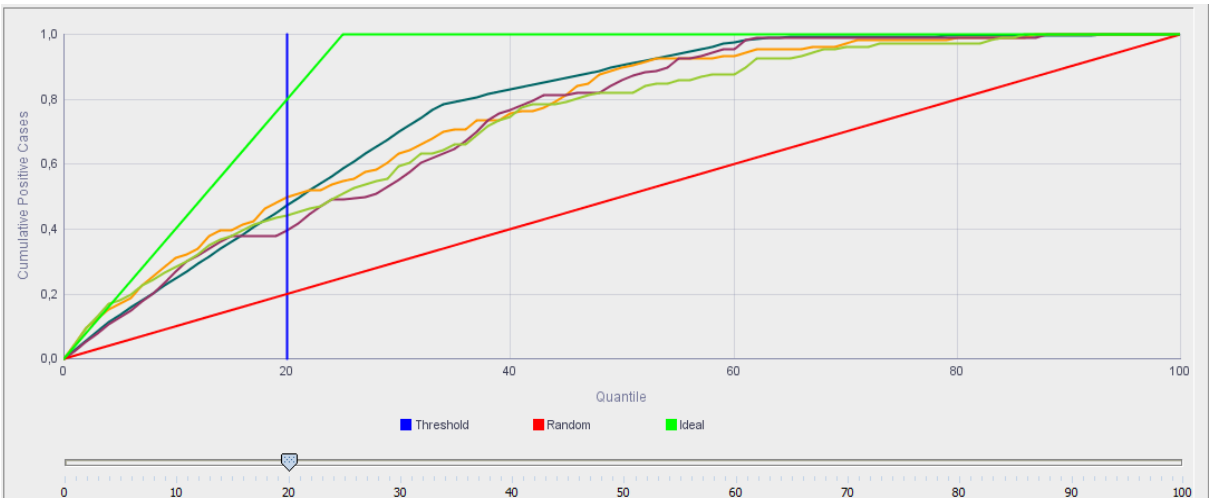


***Рис. 2. Дані про побудовані моделі в Oracle Data Miner***

Усі алгоритми, які використовувались для класифікації моделі мають схожу структуру вхідних даних:

одиничний ключовий стовпець – кожна модель повинна містити один числовий або текстовий стовпець, який унікальним чином визначає кожний запис. Використання складових ключів не допускається;

вхідні стовпці – в моделі спрощеного алгоритму Баєса всі стовпці повинні бути дискретними або дискретизованими. Для моделі спрощеного алгоритму Баєса також важливо забезпечити незалежність вхідних атрибутів один від одного. Це особливо важливо, коли модель використовується для прогнозування. Причина цього полягає в тому, що якщо використовувати два стовпці даних, які тісно пов'язані між собою, то це призведе до множення їх значень, що може затруднити інтерпретацію інших впливових факторів. І навпаки, можливість алгоритму визначати зв'язки

між змінними корисна при дослідженні моделі або набору даних для виявлення зв'язків між вхідними даними;

принаймні один прогнозований стовпець – прогнозований атрибут повинен містити дискретне або дискретизоване значення.

Після обробки моделі результати зберігаються у вигляді набору закономірностей і статистики, які можна використовувати для дослідження зв'язків або для виконання прогнозів.

Таким чином вдається отримати дані про моделі всередині окремого інструменту аналізу даних для різних моделей. Проте отримані результати можна порівняти між собою. Вони показують, що, незважаючи на однакові вхідні дані, вихідні результати для розглянутих засобів відрізняються.

Для перевірки результатів застосування методів кластеризації було використано тестові дані Oracle схеми SH (Sales History), що містить записи про клієнтів, що здійснювали покупки. На виході очікувалось отримати кластери покупців, згрупованих за певними ознаками. Оскільки кількість кластерів при використанні різних алгоритмів буде різною, для можливості порівняння виділяємо 3 групи.

Алгоритм кластеризації Microsoft надає два методи створення кластерів. Перший метод, алгоритм k-середніх, – метод жорсткої кластеризації. Це означає, що точка даних може належати лише одному кластеру і для приналежності кожної точки даних цього кластеру обчислюється одне значення ймовірності. Другий метод, максимізації очікувань (EM), – це метод м'якої кластеризації. Це означає, що точка даних завжди належить до кількох кластерів і для всіх можливих поєднань точок даних з кластерами обчислюються ймовірності. Якщо в процесі формуються порожні кластери або кількість елементів в одному або декількох групах виявляється менше заданого мінімального значення, нечисленні кластери заповнюються повторно за допомогою нових точок і алгоритм EM запускається знову.

Кластеризація методом k-середніх – добре відомий метод визначення приналежності елементів кластерам за допомогою мінімізації різниці між елементами кластера і максимізації відстані між кластерами. В обох СКБД даний алгоритм дозволяє застосовувати гнучкі налаштування без необхідності редагування реалізації, за допомогою вибору різноманітних параметрів, таких як методи кластеризації, кількість атрибутів, максимальна кількість кластерів, тощо.

В Oracle Database на додаток до вище згаданих алгоритмів визначення кластерів існує алгоритм часткової кластеризації. Цей алгоритм знаходить в ортогональних проекціях простору даних, добре виділені регіони щільності, та на їх основі ітеративно будує бінарне дерево розбиття цього простору на кластери. Однією з переваг цього алгоритму є гарна масштабуємість при обробці даних із змішаними вимірами, тому він підходить для кластеризації чисельних багатовимірних просторів з великим об'ємом даних.

Результати побудови кластерної моделі для SQL Server наведені на рис. 3.

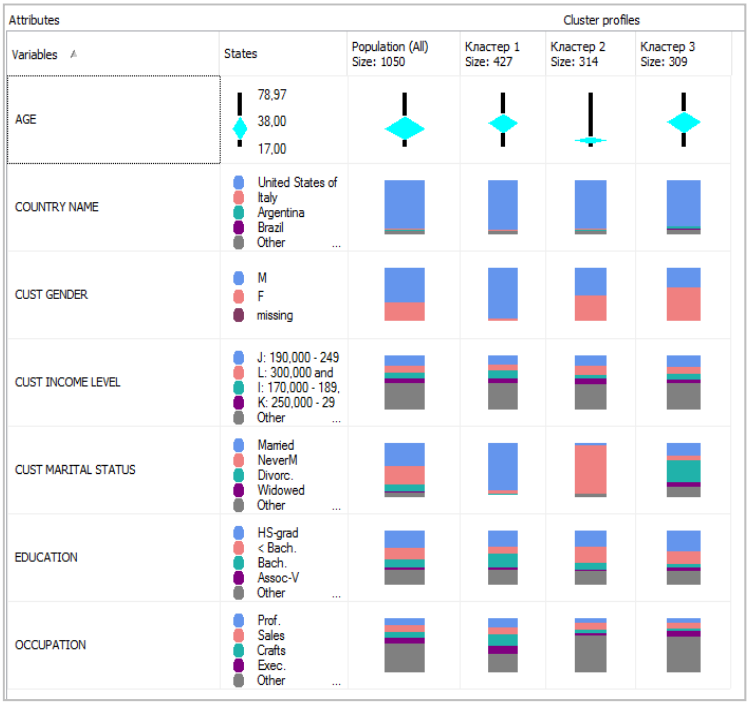


***Рис. 3. Профілі кластерів в SQL Server з гістограмами різних значень***

Оскільки, в параметрах інструменту алгоритму було вказана опція не виділяти елементи, якщо вони значно відхиляються від параметрів кластеру, тому результуюча модель містить дещо менший набір даних.

В той же час, представлення кластерів для програмних засобів Oracle базується на використанні ієрархічної моделі, тому їх представленням є дерево, листки якого і є результуючими кластерами (рис. 4).

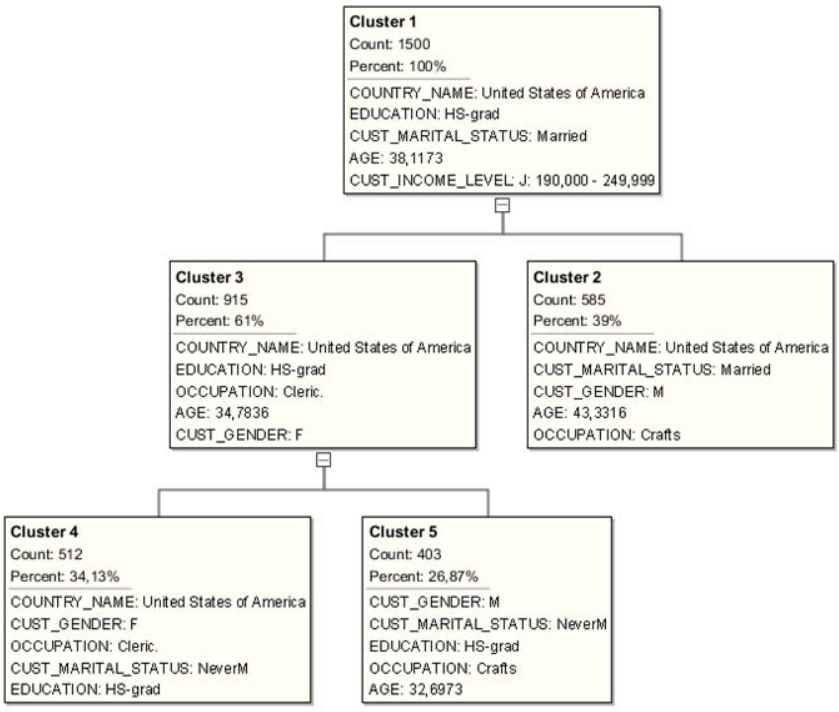


***Рис. 4. Профілі кластерів в Oracle***

В цілому результати застосування алгоритмів дають приблизно схожі результати – на тестовій вибірці було виділено три кластери, які мають практично однакові властивості та кількість входжень даних.

В обох програмних засобах для виявлення асоціацій використовується алгоритм Apriori – цей алгоритм не аналізує закономірності, замість цього він створює і підраховує потенційні набори елементів. Елемент може являти собою подію, продукт або значення атрибута, в залежності від типу аналізованих даних.

У найбільш поширеному типі моделі взаємозв'язків логічні змінні, що представляють значення «так/ні» або «існує/відсутній», присвоюються кожному атрибуту, такому як назва продукту або подія. Аналіз купівельної поведінки може слугувати прикладом моделі правил взаємозв'язку, в якій логічні змінні представляють наявність або відсутність певних продуктів в клієнтському кошику.

Потім алгоритм обчислює для кожного набору елементів рейтинг, що виражає потужність множини і достовірність. На основі цих рейтингів можна сортувати набори елементів і виводити правила.

Моделі взаємозв'язків також можуть створюватися для числових атрибутів. Якщо атрибути безперервні, числа можуть бути дискретизованої або згруповані в сегменти. Потім дискретизовані значення можна обробляти як логічні величини або як пари атрибут-значення.

На рис. 5 представлені результати отримання аналітичної моделі за допомогою методу Apriori.

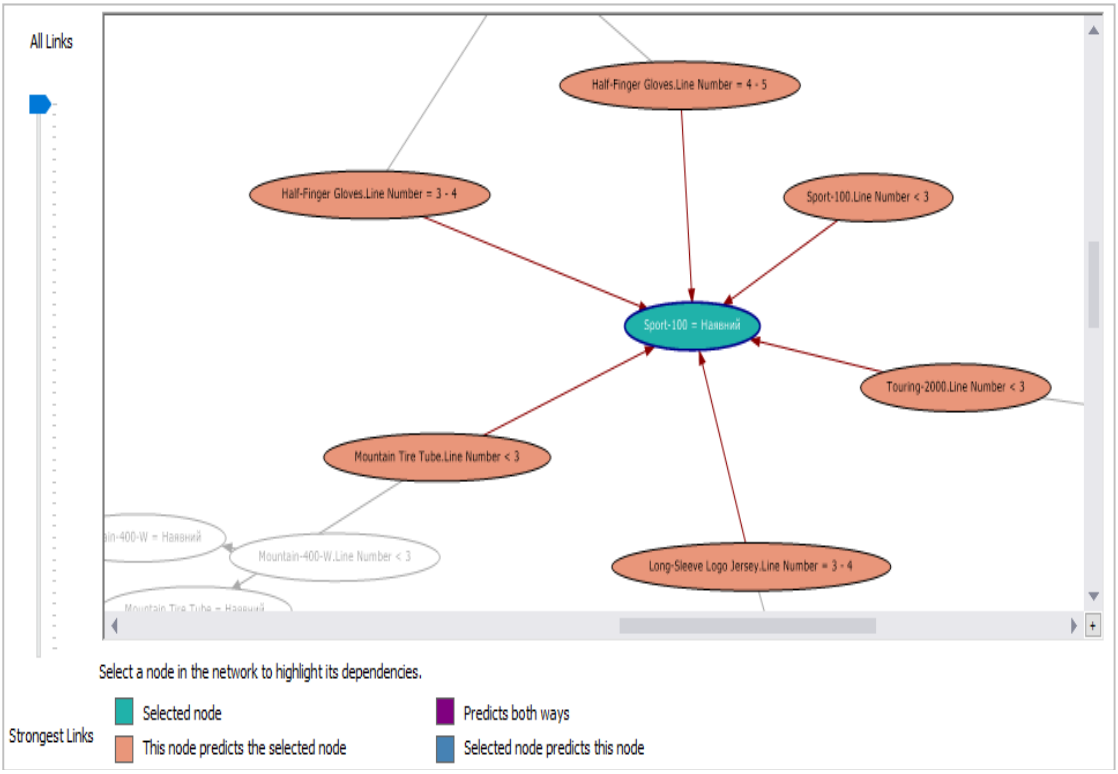


*Рис. 5. Сітка залежностей в SQL Server Data Tools*

Моделі взаємозв'язків побудовані на наборах даних, що містять ідентифікатори для окремих варіантів і елементів цих варіантів. Група елементів у варіанті

називається набором елементів. Модель взаємозв'язків складається з рядів наборів елементів і правил, що описують, як ці елементи групуються в варіантах. Правила, що визначаються алгоритмом, можуть використовуватися для прогнозування ймовірних майбутніх покупок на основі елементів, вже наявних в кошику покупця.

Алгоритм взаємозв'язків потенційно може знаходити в наборі даних багато правил. Для опису набору елементів і формованих ними правил алгоритм використовує два параметри: підтримка і ймовірність. В обох СКБД ці параметри співпадають, але в SQL Server Data Tools майстер інтелектуального аналізу не виділяє конкретних особливостей параметрів та входів, тому їх додатково доводиться вказувати у вікні налаштування алгоритму.

Для побудови моделі асоціації використовуються дані таблиць vAssocSeqOrders і vAssocSeqLineItems, що входять до складу бази даних AdventureWorks, та в яких зберігаються записи про зв'язані продажі товарів. На рис. 6. представлений потік операцій для перевірки моделі.

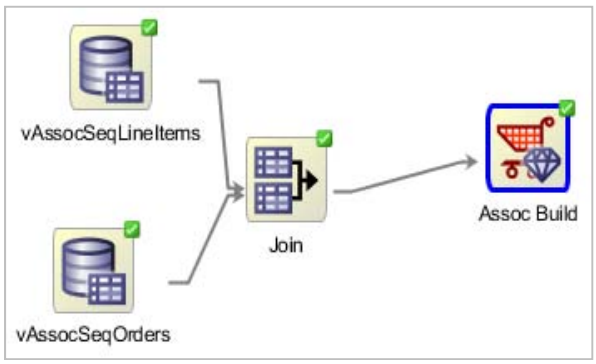


***Рис. 6. Підготовка побудови моделі асоціації в Oracle Data Miner***

Зважаючи на те, що використовувався однаковий набір даних для обох СКБД, та реалізація алгоритмів принципово не відрізняється результати виведення асоціативних правил виявились однаковими.

Слід також зазначити, що алгоритми Microsoft, а саме – алгоритм дерев прийняття рішень і алгоритм правил взаємозв'язків – можна використовувати для аналізу взаємозв'язків, але правила, виявлені кожним з алгоритмів, можуть відрізнятися. В моделі дерева прийняття рішень розбиття на вузли, що ведуть до певних правил, проводиться відповідно до отриманої інформації. У моделі взаємозв'язків правила засновані на достовірності. Тому в моделі взаємозв'язків сильне правило, або правило з високою вірогідністю, не обов'язково буде розглянуто, оскільки воно може не надавати ніякої нової інформації.

## ВИСНОВКИ

В міру того, що програмні реалізації алгоритмів та підходи до налаштувань параметрів у відповідних СКБД іноді суттєво відрізняються, результати створення аналітичних моделей можуть мати відмінності. Таке положення речей вказує на те, що конкретні алгоритми Data mining мають велику кількість варіантів реалізації, що в свою чергу ставить необхідність для користувачів чітко розуміти що саме вони планують отримати при застосуванні аналізу даних.

В цілому обидві СКБД постачають універсальні методики та широкий спектр налаштувань, що дозволяє охоплювати майже всі типові задачі аналізу даних. Також потрібно врахувати той факт, що існує можливість додання власних методів у вигляді плагінів та функцій, що в свою чергу, означає більшу гнучкість в отриманні необхідних результатів побудови аналітичних моделей.

Жоден з програмних засобів не є об'єктивно кращим, ніж інший, але деякі ситуації можуть бути більш сприятливі для певного вибору. Так, наприклад, якщо необхідна тісна інтеграція процесу розробки програмного забезпечення із функціями роботи СКБД та побудови аналітичних моделей – то кращим вибором для розробників стане SQL Server, зважаючи на доступність засобів Data mining, інтегрованих у середовище розробки та реалізації підтримки проектів з аналізу даних в програмних рішеннях.

**ПЕРЕЛІК ПОСИЛАНЬ**


1. Bassan, A. B., Sarkar, D. Mastering SQL Server 2014 Data Mining, Packt Publishing Ltd., 2014, pp. 285.
2. Zhigaylo O. M. Vikoristannya tehnologiyi data mining v avtomatizovaniy sistemi prostezhuvannosti virobnitstva siroyi sonyashnikovoyi oliyi [Elektronniy resurs] / O. M. Zhigaylo // Avtomatizatsiya tehnologichnih i biznes-protsesiv. – 2014. – № 3. – S. 30-38. – Rezhim dostupu: http://nbuv.gov.ua/UJRN/avtib_2014_3_8
3. Fisun M. T. Integratsiya tehnologiy OLAP ta Data Mining pri pobudovi mizhvimirovih asotsiativnih pravil [Elektronniy resurs] / M. T. Fisun, G. V. Gorban // ScienceRise. – 2015. – № 6(2). – S. 103-111. – Rezhim dostupu: http://nbuv.gov.ua/UJRN/texc_2015_6(2)__22
4. Shuba I. V. Ispolzovanie metodov Data Mining pri analize sotsialnyh yavleniy [Elektronniy resurs] / I. V. Shuba // Sistemi obrobki informatsiyi. – 2014. – Vip. 6. – S. 107-111. – Rezhim dostupu: http://nbuv.gov.ua/UJRN/soi_2014_6_28
5. Yuhno I. A. Primenenie tehnologiy Oracle data mining pri analize Internet-kontenta [Elektronniy resurs] / I. A. Yuhno, A. I. Yuhno // Sistemi obrobki informatsiyi. – 2009. – Vip. 7. – S. 148. – Rezhim dostupu: http://nbuv.gov.ua/UJRN/soi_2009_7_60


**Рецензент:** д.т.н., проф. Ходаков В.Е.
Херсонский национальный технический университет