УДК 681.3

Олена Дереза, кандидат технічних наук,
доцент, доцент кафедри інженерної механіки
та комп`ютерного проектування,
Галина Антонова, старший викладач кафедри
інженерної механіки та комп`ютерного
проектування,
Ілля Тетервак, асистент кафедри інженерної
механіки та комп`ютерного проектування,
Карина Валієва, здобувачка бакалаврського
рівня вищої освіти,
Таврійський державний агротехнологічний
університет імені Дмитра Моторного,
м. Запоріжжя, Україна

АНАЛІТИЧНІ ДОСЛІДЖЕННЯ МЕТОДИКИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

Анотація. Розглянуто методи інтелектуального аналізу даних. Проаналізовано сферу застосування інтелектуального аналізу даних та існуючі системи. Обговорюються відмінності DataMining від класичних статистичних методів аналізу і OLAP-систем.

Ключові слова: інтелектуальний аналіз даних, прогноз невідомих значень, прогностичне моделювання, прогнозування розвитку процесів, DataMining, OLAP-системи.

Abstract. The methods of intelligent data mining are considered. Data mining application spheres and existing systems are analyzed. Differences of data mining from classical statistical methods of analysis and OLAP-systems are discussed.

Keywords: intelligent data analysis, forecasting of unknown values, predictive modeling, process development forecasting, DataMining, OLAP systems.

Розвиток методів запису і зберігання даних привів до бурхливого зростання об'ємів збираної і аналізованої інформації. Об'єми даних настільки значні, що людині просто не під силу проаналізувати їх самостійно, хоча необхідність проведення такого аналізу цілком очевидна, адже в цих «сирих даних» укладені знання, які можуть бути використані при ухваленні рішень [1].

Алгоритми традиційної математичної статистики тривалий час, як основні, підтримували концепцію усереднення з вибірки, що приводить до операцій над

фіктивними величинами. Методи математичної статистики виявилися корисними головним чином для перевірки наперед сформульованих гіпотез і для «грубого розвідувального аналізу», що становить основу оперативної аналітичної обробки даних OLAP [2].

В основу сучасної технології DataMining встановлена концепція шаблонів, що відображають фрагменти багатоаспектних взаємостосунків в даних. Цими шаблонами є закономірності, властиві підвибіркам даних, які можуть бути компактно виражені у формі, зрозумілій людині. Пошук шаблонів проводиться методами, не обмеженими рамками апріорних припущень про структуру вибірки і вид розподілів значень аналізованих показників.

До методів і алгоритмів інтелектуального аналізу даних належать такі**:** штучні нейроні мережі, дерева рішень, методи кластерного аналізу, метод опорних векторів, метод обмеженого перебору, еволюційне програмування і генетичні алгоритми, байєсові мережі, методи пошуку асоціативних правил, кореляційно-регресійний аналіз, різноманітні методи візуалізації даних і безліч інших методів.

DataMining – міждисциплінарна галузь, що виникла і розвивалася на базі таких наук як прикладна статистика, розпізнавання образів, штучний інтелект, теорія баз даних та ін. [3]. Більшість методів інтелектуального аналізу даних була розроблена у межах теорії штучного інтелекту.

Сфера застосування DataMining нічим не обмежена – вона скрізь, де є якісь дані. Але насамперед методи DataMining сьогодні зацікавили комерційні підприємства**,** що розгортають свої проекти на основі інформаційних сховищ даних.

Метою роботи є проведення порівняльного аналізу методів інтелектуального аналізу даних та аналіз сфери застосування цих методів.

Більшість аналітичних методів, що використовуються в технології DataMining - це відомі математичні алгоритми і методи. Новим в їх застосуванні є можливість їх використання при рішенні тих або інших конкретних проблем, обумовлених новими можливостями технічних і програмних засобів, що

з'явилися.

Дейтамайнінг (Datamining) - клас аналітичного прикладного програмного забезпечення, яке підтримує рішення, розшукуючи за прихованими взірцями (patterns, шаблонами, формами, зразками, образами) інформацію в базах даних. Цей пошук може бути зроблений або за допомогою користувача (тобто тільки за допомогою виконання запитів), або інтелектуальною програмою, яка автоматично розшукує в базах даних і знаходить значущі для користувача взірці (patterns). Виконані інформаційні потреби подаються в бажаній для користувача формі, з діаграмами, звітами тощо.

Інтелектуальний дейтамайнінг відкриває інформацію всередині баз і сховищ даних, в яких користувачі не можуть ефективно виявити запити і звіти даних. Інструментальні засоби дейтамайнінгу знаходять взірці в даних і можуть навіть виводити правила з них. Ці взірці та правила потім використовуються для створення рішень і передбачення ефекту від них. Потоки даних можуть забезпечити швидкий аналіз за допомогою фокусування уваги на найбільш важливих змінних. Різке зменшення відношення вартість/продуктивність обчислювальних систем дало змогу організаціям розпочати застосування комплексних алгоритмів, які використовуються в методах дейтамайнінгу.

Єдиної думки щодо того, які задачі слід відносити до Datamining, немає. Більшість авторитетних джерел перераховує наступні: класифікація, кластеризація, прогнозування, асоціація, візуалізація, аналіз і виявлення відхилень, оцінювання, аналіз зв'язків, підведення підсумків. Розглянемо деякі з них.

Класифікація (Classification). Це найпростіша і найпоширеніша задача Data Mining. В результаті розв’язання задачі класифікації виявляються ознаки, які характеризують групи об'єктів досліджуваного набору даних - класи; по цих ознаках новий об'єкт можна віднести до того або іншого класу. Для розв’язання задачі класифікації можуть використовуватися методи: найближчого сусіда (NearestNeighbor); к-ближчого сусіда (к-Nearest Neighbor); байесові мережі (BayesianNetworks); індукція дерев рішень; нейроні мережі (neuralnet works).

Кластеризація (Clustering). Кластеризація є логічним продовженням ідеї класифікації. Ця задача складніша, особливість кластеризації полягає в тому, що класи об'єктів спочатку не визначені. Результатом кластеризації є розбиття об'єктів на групи. Прикладом методу задачі кластеризації є особливий вид нейронних мереж (карти Кохонена), що самоорганізовуються без вчителя.

Асоціація (Associations). В ході розв’язання задачі пошуку асоціативних правил відшукуються закономірності між зв'язаними подіями в наборі даних. Відмінність асоціації від двох попередніх задач DataMining: пошук закономірностей здійснюється не на основі властивостей аналізованого об'єкту, а між декількома подіями, які відбуваються одночасно. Найвідоміший алгоритм розв’язання задачі пошуку асоціативних правил - алгоритм Apriori.

Послідовність (Sequence), або послідовна асоціація (sequential association) Послідовність дозволяє знайти тимчасові закономірності між транзакціями. Задача послідовності подібна асоціації, але її метою є встановлення закономірностей не між одночасно наступаючими подіями, а між подіями, зв'язаними в часі. Цю задачу DataMining також називають задачею знаходження послідовних шаблонів (sequential pattern).Правило послідовності: після події *X* через певний час відбудеться подія *Y*.

Прогнозування (Forecasting). В результаті розв’язання задачі прогнозування на основі особливостей існуючих даних оцінюються пропущені або ж майбутні значення цільових чисельних показників. Для розв’язання таких задач широко застосовуються методи математичної статистики, нейроні мережі та ін.

Візуалізація (Visualization, GraphMining). В результаті візуалізації створюється графічний образ аналізованих даних. Для розв’язання задачі візуалізації використовуються графічні методи, що показують наявність закономірностей в даних.

Підведення підсумків (Summarization) - задача, мета якої - це опис конкретних груп об'єктів з аналізованого набору даних та інші.

Задачі DataMining, залежно від моделей, що використовуються, можуть бути дескриптивними і прогнозуючими. В результаті розв’язання описових

(descriptive) задач аналітик одержує шаблони, що описують дані, які піддаються інтерпретації. Ці задачі описують загальну концепцію аналізованих даних, визначають інформативні, підсумкові особливості даних.

Прогнозуючі (predictive) задачі ґрунтуються на аналізі даних, створенні моделі, прогнозі тенденцій або властивостей нових або невідомих даних. DataMining може складатися з таких стадій:

– виявлення закономірностей (вільний пошук);

– використання виявлених закономірностей для прогнозу невідомих значень (прогностичне моделювання);

– аналіз виключень - стадія призначена для виявлення і пояснення аномалій, знайдених в закономірностях.

Система інтелектуального аналізу даних на стадії вільного пошуку визначає шаблони, для отримання яких у системах OLAP, наприклад, аналітику необхідно обдумувати і створювати множину запитів.Тут же аналітик звільняється від такої роботи — шаблони шукає за нього система**.** Особливо корисне застосування цього підходу в надвеликих базах даних,де вловити закономірність за допомогою створення запитів доволі складно,для цього вимагається перепробувати безліч різноманітних варіантів.

Серед основних властивостей і характеристик методів DataMining розглянемо наступні: точність, масштабованість, інтерпретованість, можливість перевірки, трудомісткість, гнучкість, швидкість і популярність.

У таблиці 1 наведено порівняльну характеристику деяких поширених методів [1]. Оцінка кожної з характеристик проведена наступними категоріями, в порядку зростання: надзвичайно низька, дуже низька, низька / нейтральна, нейтральна / низька, нейтральна, нейтральна / висока, висока, дуже висока.

Важливою особливістю DataMining є не тривіальність розшукуваних шаблонів. Це означає, що знайдені шаблони повинні відображати неочевидні, несподівані регулярності в даних, складові так званих прихованих знань. Незважаючи на достатню кількість методів DataMining, пріоритет поступово зміщується у бік логічних алгоритмів пошуку в даних причинно-наслідкових

правил. За їх допомогою розв'язуються задачі прогнозування, класифікації, розпізнавання образів, сегментації баз даних, здобування з даних «схованих» знань, інтерпретації даних, установлення асоціацій в базах даних тощо. Результати таких алгоритмів ефективні й легко інтерпретуються.

Таблиця 1. Порівняльна характеристика методів DataMining

Алгоритм	Точність	Масштабованість	Інтерпретованість	Придатність до використання	Трудоємність	Різнобічність	Швидкість	Популярність, широта використання
класичні методи (лінійна регресія)	нейтральна	висока	висока / нейтральна	висока	нейтральна	нейтральна	висока	низька
нейронні мережі	висока	низька	низька	низька	нейтральна	низька	дуже низька	низька
методи візуалізації	висока	дуже низька	висока	висока	дуже висока	низька	надзвичайно низька	висока / нейтральна
дерева рішень	низька	висока	висока	висока / нейтральна	висока	висока	висока / нейтральна	висока / нейтральна
поліноміальні нейронні мережі	висока	нейтральна	низька	висока / нейтральна	нейтральна / низька	нейтральна	низька / нейтральна	нейтральна
k-найближчого сусіда	низька	дуже низька	висока / нейтральна	нейтральна	нейтральна / низька	низька	висока	низька

Системи інтелектуального аналізу даних застосовуються як масовий продукт для бізнес-додатків і як інструменти для проведення унікальних досліджень (генетика, хімія, медицина тощо). Лідери інтелектуального аналізу даних пов'язують майбутнє цих систем з використанням їх як інтелектуальних додатків, вбудованих у корпоративні сховища даних.

Список використаних джерел

1. Черняк О. І., Захарченко П. В. Інтелектуальний аналіз даних: підручник. Київ. нац. ун-т ім. Т. Шевченка. К. : Знання, 2014. 599 с.

2. Ситник В. Ф., Краснюк М. Т. Інтелектуальний аналіз даних (дейтамайнінг): навч. посібник. К : КНЕУ, 2007. 376 с.

3. Ситник В. Ф. Засоби дейтамайнінгу для аналізу бізнесових рішень. *Науково-технічна інформація*: науково-практичний журнал. 2002. №3. С. 60-64.

4. Мацулевич О. Є., Щербина В. М. Використання пакету прикладних програм NETCRACKER. *Фундаментальна підготовка фахівців у природничо-математичній, технічній, агротехнологічній та економічній галузях*: матеріали

Всеукраїнської наук.-практ. конференції з міжнар. участю, м. Мелітополь, 11-13 вересня 2017 р., присвяченої 85-річчю кафедри вищої математики і фізики, ТДАТУ. Мелітополь, 2017. С. 107-108.