УДК 330.4:519.86

Ю.М. ЛИСЕЦЬКИЙ

ТЕХНОЛОГІЇ АНАЛІЗУ ДАНИХ ПРИ ПОБУДОВІ МОДЕЛЕЙ ЕКОНОМІЧНИХ СИСТЕМ

Анотація. *Розглянуто проведення аналізу даних з використанням OLAP-технологій, Data Mining, апарату теорії нечітких множин та експертних технологій при побудові моделей економічних систем для підвищення ступеня їх адекватності.*
Ключові слова: *модель, економічна система, технологія, OLAP, Data Mining, Fuzzy sets, експертні оцінки.*

Вступ

Моделювання – один з основних інструментів досліджень у різних галузях економіки, за допомогою якого можна оцінити характеристики економічних систем для прийняття обґрунтованих управлінських рішень.

Перед тим як розпочати побудову моделі будь-якої економічної системи, необхідно провести аналіз первинних даних, які можуть бути статистичною інформацією про аналогічні системи, ретроспективними даними або оцінками експертів.

Метою цієї статті є розглядання сучасних технологій аналізу даних у контексті рішення завдання побудови моделей економічних систем заради підвищення ступеня їх адекватності.

1. Технології аналізу даних

Для проведення якісного аналізу даних пропонується використовувати OLAP-технології, Data Mining, апарат теорії нечітких множин та експертні технології.

OLAP (англ. online analytical processing – аналітична обробка у реальному часі) – технологія обробки інформації, яка дозволяє швидко отримати відповіді на багатомірні аналітичні запити. OLAP-технологія бере свій початок з 1993 року, коли засновник реляційного підходу до побудови баз даних Едгар Кодд опублікував статтю «Забезпечення OLAP для користувачів-аналітиків». У цій статті він сформулював 12 особливостей технології OLAP, які з часом було доповнено ще шістьма, і ці положення стали основним змістом нової та перспективної технології [1]. На сьогодні «OLAP» – це не тільки багатовимірний погляд на дані з точки зору кінцевого користувача, але й багатовимірне відображення їх у цільовій базі даних. Саме це привело до появи таких термінів, як ROLAP (Реляційний OLAP) і MOLAP (Багатовимірний OLAP). ROLAP-куб та система відповідних математичних алгоритмів статистичної обробки дозволяють аналізувати дані будь-якої складності на будь-яких часових інтервалах [2].

Маючи у розпорядженні гнучкі механізми маніпулювання даними та візуального відображення, дослідник спочатку розглядає з різних боків дані, що або пов'язані, або не пов'язані із проблемою, яку вирішують. Далі зіставляють різні показники між собою та намагаються виявити приховані взаємозв'язки. Після цього, за допомогою модуля статистичного оцінювання та імітаційного моделювання, будують кілька варіантів розвитку подій та обирають найбільш прийнятний варіант [1]. OLAP можна застосовувати всюди, де є завдання аналізу багатофакторних даних. Після налаштування на дані користувач має можливість швидко отримувати відповіді на ключові питання шляхом простих маніпуляцій мишею над OLAP-таблицею та відповідними меню. При цьому будуть доступні певні стандартні методи аналізу, котрі логічно випливають з природи OLAP-технології: факторний (структурний) аналіз, аналіз динаміки (регресійний аналіз – знаходження трендів), аналіз залежностей (кореляційний аналіз), порівняльний аналіз, дисперсійний аналіз (дослідження розподілення ймовірностей та довірчих інтервалів показників, що розглядаються). Цими видами аналізу можливості OLAP не вичерпуються. Наприклад, якщо використати як алгоритм обчислення проміжних та кінцевих підсумків функції статистичного аналізу дисперсію, середнє відхилення, моди більш високих порядків, можна отримати більш детальні види аналітичних звітів [2].

Таким чином, OLAP-технологія є інструментом для аналізу великих обсягів даних у режимі реального часу. За допомогою OLAP дослідник може здійснити гнучкий перегляд інформації, отримати довільні зрізи даних та виконати аналітичні операції деталізації, згортки, наскрізного розподілення, одночасні порівняння у часі за багатьма параметрами. Програмні засоби OLAP – це інструмент оперативного аналізу даних, головна особливість яких – це орієнтація на використання не IT-фахівцем, не експертом-статистиком, а професіоналом у прикладній галузі. Вся робота з OLAP відбувається у термінах предметної галузі та дозволяє будувати статистично обґрунтовані моделі [2].

Data Mining (видобування даних, інтелектуальний аналіз даних, глибинний аналіз даних) – це збиральна назва, яку використовують для позначення сукупності методів, що дозволяють виявляти знання у раніше відомих базах даних. Термін було введено Г. Пятецьким-Шапіро у 1989 р. Він є поєднанням широкого математичного інструментарію та останніх досягнень у сфері інформаційних технологій [3]. У технології об'єднані строго формалізовані методи та методи неформального аналізу даних. Основу методів Data Mining складають методи класифікації, моделювання та прогнозування. Одне з найважливіших призначень методів Data Mining полягає у наочному поданні результатів обчислень, що дозволяє використовувати інструментарій Data Mining особам, які не мають спеціальної математичної підготовки. Знання, що здобувають методами Data Mining, зазвичай подають у вигляді моделей (рис. 1).

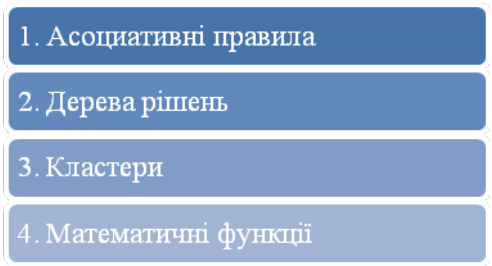

Рисунок 1 – Моделі подання знань Data Mining

Методи та алгоритми побудови таких моделей, зазвичай, відносять до галузі штучного інтелекту, так як більшість з них було розроблено у межах теорії штучного інтелекту.

Таким чином, до методів та алгоритмів Data Mining відносяться [3]: штучні нейронні мережі; дерева рішень, символьні правила; методи найближчого сусіда та k-ближчого сусіда; методи опорних векторів; байєсівські мережі; лінійна регресія; кореляційно-регресійний аналіз; ієрархічні методи кластерного аналізу, у тому числі алгоритми k-середніх та k-медіани; методи пошуку асоціативних правил, у тому числі алгоритм Apriori; метод обмеженого перебору; еволюційне програмування та генетичні алгоритми; різноманітні методи візуалізації даних та багато інших методів. Більшість аналітичних методів, що використовують у технології Data Mining – це відомі математичні алгоритми та методи. Новою у їх застосуванні є можливість використати їх під час рішення тих чи інших конкретних проблем, що обумовлена виниклими можливостями технічних та програмних засобів.

Fuzzy sets – побудова математичної моделі за результатами спостереження або завдання ідентифікації систем [4]. Використовується у тих випадках, коли моделі, що синтезують, базуються на експертних лінгвістичних висловлюваннях. Одним з найбільш розроблених в інженерному відношенні інструментів обліку лінгвістичної інформації є теорія нечітких множин та нечітка логіка, яка бере свій початок з 1965 р., коли професор Лотфі Заде з Каліфорнійського університету Берклі опублікував основоположну статтю «Fuzzy Sets» у журналі «Information and Control» [5].

Практично завжди побудова аналітичної системи аналізу даних – це завдання побудови єдиної інтегрованої інформаційної системи, на основі неоднорідних програмних засобів, що функціонує узгоджено [2] (рис. 2).

Якщо у якості даних використовують експертні оцінки, необхідно провести аналіз узгодженості суджень експертів для перевірки вірогідності експертних оцінок та виявлення причин їх неоднорідності [6]. Це можна зробити за допомогою статистичної обробки інформації, отриманої від експертів [7]. У цьому випадку отримані від експертів оцінки можна розглядати як випадкові змінні і тому для аналізу розкиду узгодженості оцінок використовують наступні статистичні характеристики [8]:

– середнє значення оцінок (точкова оцінка для даної групи експертів), яке характеризує узагальнене судження експертів щодо альтернатив;

– середнє квадратичне відхилення, що характеризує розкид суджень окремих експертів відносно середнього значення;

– коефіцієнт варіації, що характеризує варіабельність, яку розраховують у вигляді відношення середнього квадратичного відхилення оцінки до середньої арифметичної.

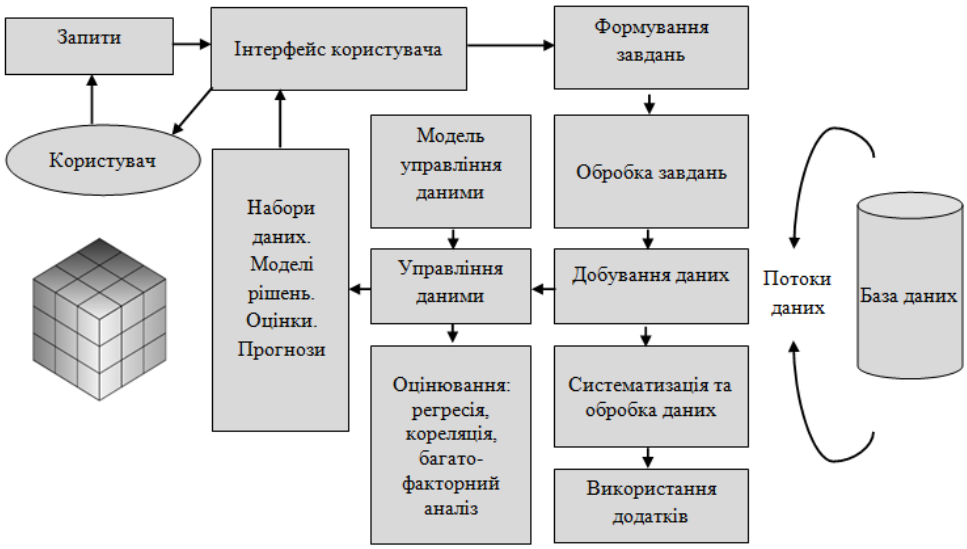

Рисунок 2 – Інформаційно-аналітична система вилучення, отримання і обробки даних

З точки зору математичної статистики, оцінки, що суттєво відрізняються від середнього значення, можна вважати випадковими. Тому було введено поняття суперечливості судження експерта k узагальненому судженню всіх експертів. Воно базується на припущенні, що судження y_k експерта k є крайнім серед суджень m експертів. Аналіз суперечливості судження експерта k проводять з використанням оцінки анормальності результатів при невідомій генеральній дисперсії [6].

Для оцінки ступеня подібності суджень експертів використовують коефіцієнти асоціації (за Устюжаніновим), за допомогою яких враховують лише кількість відповідей, що співпадають або не співпадають, та не враховують їх послідовність [8].

Для більш точної оцінки узгодженості суджень експертів використовують методи рангової кореляції:

1. Коефіцієнт рангової кореляції Кендала як одну з вибіркових мір залежності двох випадкових величин (ознак) X та Y, що заснована на ранжуванні елементів вибірки $(X_1, Y_1),...,(X_n, Y_n)$. Коефіцієнт рангової кореляції Кендала відноситься до рангових статистик, та як будь-яку рангову статистику його можна використовувати для знаходження залежності двох якісних ознак, тільки якщо елементи вибірки можна упорядкувати відносно цих ознак [6].

2. Коефіцієнт рангової кореляції Спірмена, який теж є мірою залежності двох випадкових величин, засновано на ранжуванні незалежних результатів спостережень, за допомогою яких коефіцієнт можна обчислити простіше та швидше [6].

Під час аналізу оцінок, отриманих від експертів, часто виникає необхідність виявити узгодженість їх суджень щодо декількох альтернатив, що впливає на один кінцевий результат. У цьому випадку узгодженість суджень експертів можна оцінити за допомогою коефіцієнта конкордації – загального коефіцієнта рангової кореляції для групи, що складається з m експертів [6].

Для того щоб оцінити значущість коефіцієнта конкордації, використовують критерій χ^2. Знайдене значення повинно бути більше табличного значення χ^2, що визначається кількістю ступенів його свободи та рівнем довірчої вірогідності. Це підтверджує значущість коефіцієнта конкордації.

Для прискорення визначення узгодженості та вірогідності експертних оцінок розроблено технологію аналізу експертних суджень шляхом використання послідовності методів знаходження їх неоднорідності. Сутність технології полягає у наступному [6].

У першу чергу необхідно оцінити узгодженість суджень експертів за допомогою коефіцієнта конкордації. Якщо коефіцієнт конкордації є значущим, то судження групи експертів узгоджені та подальший аналіз можна не проводити.

Якщо судження експертів виявляються неузгодженими, то для оцінки ступеня подібності суджень кожної пари експертів треба розрахувати коефіцієнти асоціації за Устюжаніновим.

Для більш точної перевірки узгодженості суджень експертів необхідно використовувати метод рангової кореляції Спірмена, за допомогою якого коефіцієнт можна обчислити легше та швидше, ніж коефіцієнт рангової кореляції Кендала.

При наявності неузгодженості суджень експертів необхідно продовжити аналіз для знаходження причин їх неузгодженості та провести перевірку на суперечливість суджень, у ході якої знаходять експертів, чиї судження істотно відрізняються від узагальненого судження групи.

Запропонована технологія аналізу експертних висновків шляхом застосування послідовності методів знаходження неоднорідності суджень експертів може бути алгоритмізована, та алгоритм містить наступні кроки:

Крок 1. Обчислення коефіцієнта конкордації.

Крок 2. Оцінка значущості коефіцієнта конкордації.

Крок 3. Обчислення коефіцієнтів асоціації.

Крок 4. Оцінка мір подібності суджень пар експертів.

Крок 5. Обчислення коефіцієнтів рангової кореляції.

Крок 6. Оцінка узгодженості суджень експертів.

Крок 7. Обчислення узагальненого судження групи експертів.

Крок 8. Перевірка судження експерта на суперечливість узагальненому судженню групи.

Цей алгоритм програмно реалізовано на мові С++, та за допомогою програмної реалізації було проведено практичні дослідження, котрі експериментально підтвердили ефективність запропонованої технології.

Висновки

Таким чином, з урахуванням викладеного вище очевидно, що відповідальним етапом побудови моделей економічних систем, що підвищує ступінь їх адекватності, є аналіз вихідних даних, який передбачає перевірку даних, забезпечення їх порівнянності, узгодженості та вірогідності. Незважаючи на те, що аналіз даних є достатньо трудомістким процесом, його якісне проведення можливе з використанням розглянутих технологій.

СПИСОК ЛІТЕРАТУРИ

1. Технологія OLAP [Електронний ресурс]. – Режим доступу: https://studopedia.org/8-3296.html
2. OLAP-технології [Електронний ресурс]. – Режим доступу: http://oplib.ru/random/view/317374
3. Методи Data Mining [Електронний ресурс]. – Режим доступу: http://uadoc.zavantag.com/text/26515/index-1.html
4. Zadeh L. A. Fuzzy Sets, Information and Control / L. A. Zadeh // Fuzzy sets and systems / J. Fox Ed. – 1965. – N 8. – P. 338–353.
5. Історія виникнення теорії нечітких множин [Електронний ресурс]. – Режим доступу: http://um.co.ua/3/3-6/3-62583.html
6. Лисецький Ю.М. Інформаційні технології підтримки прийняття рішень при побудові корпоративних інтегрованих інформаційних систем: автореф. дис. док. техн. наук: спец. 05.13.06 «Інформаційні технології» / Ю.М. Лисецький. – К., 2017. – 39 с.
7. Литвак Б. Г. Экспертная информация: Методы получения и анализа / Литвак Б. Г. – М.: Радио и связь, 1984. – С. 118.
8. Бешелев С. Д. Математико-статистические методы экспертных оценок / С. Д. Бешелев, Ф. Г. Гурвич. – М.: Статистика, 1980. – 263 с.

Стаття надійшла до редакції 08.07.2018.