УДК 004.94

О. С. Гайденко
(аспірант кафедри «Автоматизація та комп'ютерно-інтегровані технології транспорту», Державний економіко-технологічний університет транспорту, м. Київ)

СУЧАСНІ ТЕНДЕНЦІЇ ТА ОСНОВНІ МЕТОДИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ПЕРВИННИХ ДАНИХ

Розглянуто фундаментальні методи інтелектуального аналізу первинних даних, такі як графічні та статистичні методи, штучний інтелект, дерево рішень і генетичний алгоритм. Досліджено області їхнього можливого застосування у вирішенні завдань інтелектуалізації системи електропостачання залізниці.

Ключові слова: *інтелектуальна обробка інформації, data mining, електропостачання, залізниця.*

Рассмотрены фундаментальные методы интеллектуального анализа первичных данных, такие как графические и статистические методы, искусственный интеллект, дерево решений и генетический алгоритм. Исследованы области их возможного применения в решении задач интеллектуализации системы электроснабжения железной дороги.

Ключевые слова: *интеллектуальная обработка информации, data mining, электроснабжение, железная дорога.*

Постановка проблеми. Стрімке зростання інформації у базах даних привело до необхідності розробки нових технологій та інструментів для інтелектуальної та автоматичної обробки даних в корисну інформацію та знання [1].

Для вирішення деяких завдань інтелектуальної енергетики потрібна така обробка даних, яка дозволить на підставі знайдених взаємозв'язків будувати моделі, здатні описати особливості функціонування реальних систем електропостачання [2].

Аналіз останніх досліджень і публікацій свідчить, що напрям інтелектуального аналізу первинних даних сьогодні актуальний як у діяльності, пов'язаній безпосередньо з інформацією, так і в промисловості [1-12]. Причиною цього є розвиток комп'ютерних обчислювальних систем і сховищ даних [2].

Основна частина. Первинні дані (початкові, сирі або історичні дані) – необроблені масиви даних, отримувані в результаті спостереження за динамічною системою або об'єктом, що відображають його стан у конкретні моменти часу.

У випадку системи електропостачання залізниці джерелом первинних даних є пристрої моніторингу – різного роду датчики та елементи системи комерційного обліку АСКОЕ.

Оброблені дані несуть у собі інформаційну цінність, тобто є вже не даними, а інформацією.

Також у сучасному інтелектуальному аналізі даних data mining (DM) популярним є термін «видобуток знань» із даних. Знання в data mining – це знайдені взаємозв'язки між об'єктами аналізу.

Розглянемо всі сучасні фундаментальні підходи DM, які можуть використовуватись для завдань інтелектуалізації системи електропостачання залізниці.

Статистичні методи

Статистичні методи є незамінним компонентом у відборі та аналізі даних, і оцінці здобутих знань. Вони використовуються для оцінки результатів DM, щоб відокремити корисні дані від негативних. Статистичні методи пропонують використовувати у процесі очищення даних, для виявлення «сторонньої» інформації, а також для оцінки спотворень. Статистичні методи також можуть використовуватися для роботи з відсутніми даними, використовуючи засоби оцінки.

Методи для транзакцій і реляційної бази даних

Як відомо, транзакція – логічна одиниця роботи з даними, що являє собою окрему послідовність операцій із БД. Інтелектуальний пошук правил асоціацій популярний для використання у транзакціях БД і реляційних БД. Завдання полягає в тому, щоб отримати набір чітких асоціативних правил у формі

$$X_1 \dots X_n \Rightarrow Y_1 \dots Y_{m,}$$

де X_i (для $i \in 1,\dots,n$) та Y_j (для $j \in 1,\dots,m$) – атрибутизначень з множини відповідних наборів інформації у базі даних. Наприклад, у одній і тій самій транзакції можна знайти правило про асоціацію: якщо потяг спізнюється відносно розкладу, то він зазвичай витрачає більше енергії на проходження тієї самої ділянки. Оскільки операції DM для транзакцій бази даних можуть вимагати повторюваного читання (Repeatable Read, Snapshot), може знадобитися величезна потужність обробки [3].

Методи штучного інтелекту (Artificial intelligence (AI))

Методи AI широко застосовуються в DM. Методам розпізнавання образів, машинного навчання і нейронних мереж приділяється велика увага. Інші методи в AI, такі як здобуття знань, представлення знань і пошуку, мають відношення до різних етапів процесу в DM.

Одним із рішень класифікації як одного з основних завдань DM є використання нейронної мережі. Класифікація – це процес розподілу набору даних на взаємовиключні групи.

Нейронна мережа для DM складається з трьох основних етапів [4]:

1. Будівництво мережі та навчання: в цій фазі багаторівнева нейронна мережа створюється та навчається на основі заданих атрибутів, класів та обраного методу.
2. Спрощення мережі: на даному етапі надлишкові зв’язки та блоки видаляються без збільшення частоти помилок класифікації у мережі.
3. Вилучення правил класифікації.

Генетичний алгоритм

Генетичний алгоритм – відносно нова парадигма програмного забезпечення, підґрунтям якого є теорія еволюції Дарвіна. Популяція правил, кожна з яких відображає можливе рішення проблеми у вигляді хромосом, спочатку створюється випадковим чином. Тоді пари правил (як правило, найсильніші правила вибираються як «батьки») об’єднуються, щоб зробити «потомство» для наступного покоління. Нова популяція рішень формується застосуванням двох генетичних операторів. Перший – секції хромосом двох рішень міняються місцями та дають нові рішення. Другий – використовується процес мутації для зміни випадковим чином генетичної структури деяких членів кожного нового покоління. Система працює протягом кількох десятків або сотень поколінь. Процес завершується, коли буде виконана наперед задана умова (прийнятне або оптимальне рішення знайдене, або після деякої межі, встановленого часу) [5]. Ге-

нетичні алгоритми підходять для вирішення завдань, що вимагають оптимізації за певним обчислюваним критерієм. Ця парадигма може бути застосована до завдань DM. Величина, яка повинна бути зведена до мінімуму, часто є числом помилок класифікації на навчальній вибірці. Щоб отримати відповідні результати в розумні терміни, великі та складні завдання вимагають потужного апаратного забезпечення. Видобуток великих масивів даних за допомогою генетичних алгоритмів набирає практичної популярності лише останнім часом у зв'язку з наявністю доступних високошвидкісних комп'ютерів.

Графічні методи DM

Візуальні методи DM довели цінність пошукового аналізу даних, і вони також мають хороший потенціал для роботи з великою базою даних. Такий підхід вимагає інтеграції людини в процес DM. Є кілька добре відомих методів для візуалізації багатовимірних масивів даних: діаграми, графіки, паралельні координати, матриці проекцій, а також інші методи геометричної проекції, такі як ієрархічні методи, засновані на графах.

Часто графічні моделі передають взаємозв'язки з використанням структури графа [6,7]. У своїй простій формі, модель визначає, які змінні безпосередньо залежать одна від одної.

Дерева рішень

Дерева рішень – набори рішень у деревовидній формі для кращого сприйняття людиною. Вони в основному використовуються для прогнозного моделювання, класифікації набору даних по заданих правилах і завдань регресійного аналізу [8].

Для ухвалення рішення слід дати відповідь на питання виду «значення Y<x?» у вузлах дерева, починаючи від кореня.

Конкретні методи дерев рішень у DM, бувають двох основних типів [9,10]:

- класифікації, результатом роботи яких є присвоєння категорії для даних;
- регресійного аналізу, коли результат можна сприймати як дійсне число (наприклад, час руху потяга або вартість спожитої електроенергії).

Classification and Regression Trees (CART) і Chi Square Automatic Interaction Detection (CHAID) – популярні алгоритми дерев рішень, які використовуються для класифікації набору даних. Вони визначають набір правил, які можуть бути застосовані до нового (несортованого) набору даних для прогнозування.

CART, як правило, вимагає менше підготовки даних, ніж CHAID. CART може обробляти відсутні значення. Побудована модель може бути перевірена на окремо зазначеному тестовому наборі даних, крім того, вона може бути збережена і використана згодом на додаткових тестових наборах [9].

Велика кількість алгоритмів дерев рішень описано в літературі з області машинного навчання і прикладної статистики [11,12].

Гібридні методи

Очевидно, що гібридні методи в інтелектуальному аналізі не належать до основних, проте через зростання їхньої популярності проігнорувати їх було б недоречно.

Ідея алгоритмів побудованих на основі гібридних методів полягає в комбінації декількох методів для знаходження наближеного рішення без гарантій якості. Головною особливістю таких систем є використання окремого модуля для прийняття рішень, в який, наприклад, надходить інформація з нейронної мережі і структуровані дані з експертної системи. Використання таких систем виправдане при автоматизації аналітичних процесів обробки великого масиву даних з подальшою агрегацією, а також при автоматизації складних технологічних процесів. Однією з останніх тенденцій у розробці гібридних інтелектуальних алгоритмів і систем є об'єднання класичних методів AI з нечіткими системами, генетичними алгоритмами і т. п. Проте при впровадженні гіб-

ридних інтелектуальних систем часто виникають проблеми, пов'язані з інтеграцією окремих компонентів, через складнішу архітектуру такі системи мають низьку відмовостійкість [5].

Висновки. Наведені підходи інтелектуального аналізу первинних даних дозволяють вирішувати актуальні завдання, такі як прогнозування електроспоживання, зміни графіка руху потягів, ефективної діагностики та керування системою електропостачання залізниці тощо. На підставі розглянутих методів при практичному застосуванні можуть виникати нові нетрадиційні алгоритми, пристосовані до виконання конкретних завдань.

ЛІТЕРАТУРА

1.*Sang Jun Lee* A review of data mining techniques / *Sang Jun Lee, Keng Siau* // Industrial Management & Data Systems, Vol. 101 Iss: 1, pp.41–46

2.О.С. Гайденк. Інтелектуальна обробка баз знань господарства електропостачання залізниць // Збірник наукових праць ДЕТУТ. Серія «Транспортні системи і технології». – Вип. 28. – К.: ДЕТУТ, 2016. – С. 147-153.

3.Jim Gray. The Transaction Concept: Virtues and Limitations. Proceedings of the 7th International Conference on Very Large Databases, 1981, pp. 144–154.

4.Lu H. Effective Data Mining Using Neural Networks / Lu H., Setiono R., Liu H. // IEEE Transactions on Knowledge and Data Engineering – 1996, Vol. 8 No. 6. – pp. 957-961.

5.В.А. Гречкин. Интеллектуальные алгоритмы обработки информации в многокритериальных системах поддержки принятия решений // Вестник Ставропольского государственного университета – 2010. – №70. – С. 35-39.

6.Whittaker J. Graphical Models in Applied Multivariate Statistics // New York: Wiley – 1990.

7.Pearl J. Probabilistic Reasoning in Intelligent Systems // San Francisco, Calif.: Morgan Kaufmann – 1988.

8.Fayyad U.M. From Digitized Images to On-Line Catalogs: Data Mining a Sky Survey / Fayyad U.M., Djorgovski S.G., Weir N. // AI Magazine – 1996. – №17(2). – pp. 51–66.

9.Joshua Gould Classification and Regression Trees (CART) Documentation [Електрон. ресурс]. – Режим доступу: ftp://ftp.broad.mit.edu/pub/genepattern/modules/CART/broad.mit.edu:cancer.software.genepattern.module.analysis/00056/1/CART.pdf.

10. Breiman L. Classification and regression trees. Monterey / Breiman L., Friedman J.H., Olshen R.A., Stone C.J. // CA: Wadsworth & Brooks/Cole Advanced Books &Software, 1984.

11. Breiman L. Classification and Regression Trees / Breiman L., Friedman J.H., Olshen R.A., Stone C.J. // Belmont, Calif.: Wadsworth, 1984.

12. Quinlan J. C4.5: Programs for Machine Learning / San Francisco, Calif.: Morgan Kaufmann, 1992.

Oles Haidenko

(Graduate Student of Automation and Computer-Integrated Technologies of Transport Chair, State University for Transport Economy and Technologies)

CURRENT TRENDS AND BASIC METHODS OF RAW DATA INTELLIGENT ANALYSIS

Fundamental methods of the data mining, such as visualization and statistical methods, artificial intelligence, decision trees and genetic algorithm are considered. Fields of their possible application in the task of railroad power supply system intellectualization are research.

Keywords: *intelligent information processing, data mining, electricity, railways.*

REFERENCES

1. *Sang Jun Lee.* A review of data mining techniques / *Sang Jun Lee, Keng Siau* // Industrial Management & Data Systems, Vol. 101 Iss: 1, pp.41–46

2. O. Haidenko. Intelektualna obrobka baz znan hospodarstva elektropostachannya zaliznyc [Intelligent processing knowledge bases of railway power facilities] // Zbirnyk naukovyh prac' DETUT: Serija «Transportni systemy i tehnologii'», vol. 28. – K.: 2016. – pp. 147-153.

3.Jim Gray. The Transaction Concept: Virtues and Limitations. Proceedings of the 7th International Conference on Very Large Databases, 1981, pp. 144–154.

4.Lu H. Effective Data Mining Using Neural Networks / Lu H., Setiono R., Liu H. // IEEE Transactions on Knowledge and Data Engineering – 1996, Vol. 8 No. 6. – pp. 957-961.

5.V. Hrechkin. Intelektualnye algoritmy obrabotki informacii v mnogokriterialnyh sistemah podderjki prinyatiya resheniy [Intelligent information processing algorithms in multiobjective decision support systems] // Vestnik Stavropolskogo gosudarstvennogo universiteta – 2010. – №70. – pp. 35-39.

6.Whittaker J. Graphical Models in Applied Multivariate Statistics // New York: Wiley – 1990.

7.Pearl J. Probabilistic Reasoning in Intelligent Systems // San Francisco, Calif.: Morgan Kaufmann – 1988.

8.Fayyad U.M. From Digitized Images to On-Line Catalogs: Data Mining a Sky Survey / Fayyad U.M., Djorgovski S.G., Weir N. // AI Magazine – 1996. – №17(2). – pp. 51–66.

9.Joshua Gould. Classification and Regression Trees (CART) Documentation. ftp://ftp.broad.mit.edu/pub/genepattern/modules/CART/broad.mit.edu:cancer.software.genepattern.module.analysis/00056/1/CART.pdf

10. Breiman L. Classification and regression trees. Monterey / Breiman L., Friedman J.H., Olshen R.A., Stone C.J. // CA: Wadsworth & Brooks/Cole Advanced Books &Software, 1984.

11. Breiman L. Classification and Regression Trees / Breiman L., Friedman J.H., Olshen R.A., Stone C.J. // Belmont, Calif.: Wadsworth, 1984.

12. Quinlan J. C4.5: Programs for Machine Learning / San Francisco, Calif.: Morgan Kaufmann, 1992.