УДК 004.75

В.В. Мазуренко
Н.В. Добровольська

МЕТОДИ МАШИННОГО НАВЧАННЯ ДЛЯ ПРОГНОЗУВАННЯ У ДАТА-МАЙНІНГУ

Вінницький національний технічний університет

Анотація

У роботі узагальнено питання виявлення нової та потенційно корисної інформації з обширних обсягів даних, що підкреслює важливість розвитку інструментів інтелектуального аналізу даних для комплексних соціально-економічних процесів і систем, спрямованих на принципи цифрової економіки, та їх обробки за допомогою мережевих додатків. Описано етапи інтелектуального аналізу даних, які враховують складні соціально-економічні процеси і системи. Проаналізовано цикл обробки даних, що включає послідовні кроки від введення необроблених даних до отримання корисної інформації. Знання, отримані на етапі обробки даних, слугують основою для створення моделей складних соціально-економічних процесів і систем. Розрізняють два типи моделей (описові та прогнозні), які можуть бути розроблені в рамках інтелектуального аналізу даних.

Ключові слова: data mining, прогнозне моделювання, нейронні мережі, глибоке навчання.

Abstract

The article summarizes the issue of identifying new and potentially useful information from large data sets, which emphasizes the importance of developing data mining tools for complex socio-economic processes and systems based on the principles of the digital economy and their processing using network applications. The stages of data mining that take into account complex socio-economic processes and systems are described. The author analyzes the data processing cycle, which includes successive steps from entering raw data to obtaining useful information. The knowledge gained at the data processing stage serves as the basis for creating models of complex socio-economic processes and systems. There are two types of models (descriptive and predictive) that can be developed as part of data mining.

Keywords: *data mining, predictive modeling, neural networks, deep learning.*

Вступ

Постійне масштабування даних в Інтернеті змінює спосіб нашої взаємодії в економічних та соціальних системах. Багато користувачів щодня шукають, публікують і створюють нові дані, залишаючи цифровий слід, який може допомогти описати їхню поведінку, рішення та наміри. Це підкреслює роль розробки інструментів інтелектуального аналізу даних для складних соціально-економічних процесів і систем, заснованих на принципах цифрової економіки, та їх обробки за допомогою мережевих додатків.

Метою роботи є дослідження процесу виявлення нової та потенційно корисної інформації з великих масивів даних, окреслення етапів інтелектуального аналізу даних для складних соціально-економічних процесів і систем та визначення відповідного інструментарію в умовах прогресу обчислювальних потужностей та появи великої кількості багатовимірних даних у вільному доступі.

Методи машинного навчання для прогнозування у дата-майнінгу

Оскільки інтелектуальний аналіз даних еволюціонував як професійна діяльність, необхідно відрізняти його від попередніх статистичного моделювання та більш широкої діяльності з виявлення знань.

Інтелектуальний аналіз даних (Data визначається як використання алгоритмів машинного навчання для пошуку слабких зв'язків між елементами даних у великих і невпорядкованих наборах даних, що може привести до дій, спрямованих на збільшення вигоди в тій чи іншій формі (діагностика, прибуток, прогнозування, управління тощо) [1].

Інтелектуальний аналіз даних також називають виявленням знань у базах даних (Knowledge Discovery in Databases - KDD), тобто процес виявлення нової та потенційно корисної інформації з великих обсягів даних. Визначення інтелектуального аналізу даних спочатку обмежувалося процесом моделювання, але з часом інструменти аналізу даних аналізу даних стали включати процеси, що полегшують підготовку даних, а також оцінювання та візуалізацію моделей [2] (рис. 1).

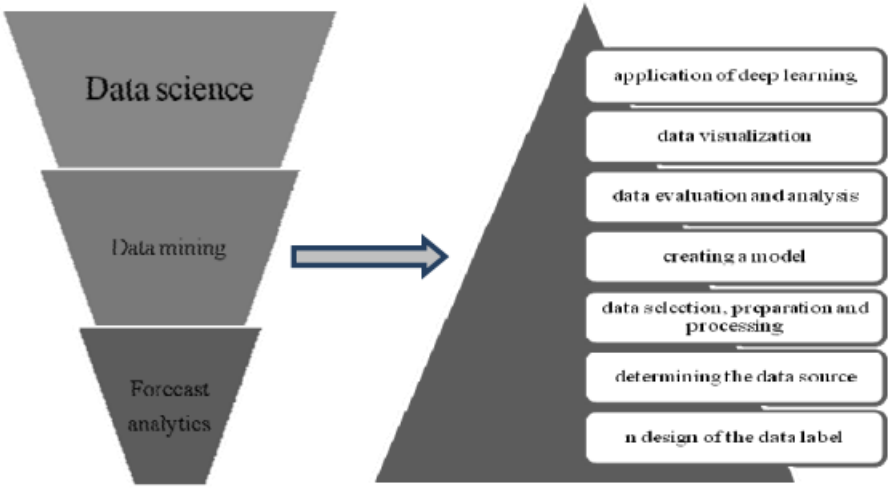

Рисунок 1 – Алгоритм аналізу даних та інструменти інтелектуального аналізу даних

Оцінка та аналіз даних. Метою будь-якого прогнозного моделювання є застосувати модель до нових даних. Прогностичні моделі корисні лише настільки, наскільки якість їхнього передбачення є адекватною, тому принциповим є не процес створення моделі як такий, а створення якісної створення моделі як такої, а створення якісної моделі. Як прогностичні, так і дескриптивні та описові моделі мають свої критерії оцінки. Для прогнозних моделей критерієм оцінки є точність прогнозу, яка вимірюється величиною помилкою прогнозу, тобто різницею між прогнозом і фактичним значенням досліджуваного показника. Для дескриптивних моделей складніше визначити очевидні критерії оцінки, але вони зазвичай фіксують розбіжність між спостережуваними даними та запропонованою моделлю. Таким чином, на цьому етапі інтелектуального аналізу даних можна використовувати різні стратегії оцінки якості моделей.

Параметричні методи аналізу точності прогнозів. Відповідно до результатами ex-post-прогнозу, такі показники точності прогнозу на m кроків, як середньоквадратична похибка кроків розраховуються такі показники точності прогнозу, як середня квадратична похибка, корінь стандартної похибки, середня абсолютна похибка, корінь середньоквадратичної похибки у відсотках, середня абсолютна відносна похибка у відсотках (MARE). Чим менше значення цих величин, тим вища якість прогнозу. На практиці ці характеристики використовуються досить часто. Такий підхід дає хороші результати, якщо в період ретро-прогнозу не з'являються принципово нові закономірності. Для створення прогностичної моделі складних соціально-економічних систем і процесів прогноз щоразу будується в новій ситуації, тому порівняння числової точності прогнозів, зроблених у різні моменти часу є не зовсім коректним. Ці міркування зумовили використання непараметричних методів аналізу точності прогнозів [2].

Непараметричні методи аналізу точності прогнозу мають два типи непараметричних критеріїв: критерій міток та критерій рангів. Критерій міток для порівняння точності двох послідовностей прогнозів базується на відсотку випадків, коли метод визначення прогнозу А є кращим за метод В. Таке порівняння проводиться для окремих прогнозів одних і тих самих подій (змінних). Якщо застосовуються ранги їхніх критеріїв, то числова характеристика точності (абсолютна похибка при оцінці одного прогнозу, або середньоквадратична похибка при розгляді послідовності прогнозів) замінюється на ранги, які потім

перевіряються на значущість. Наприклад, якщо послідовності прогнозів показників A і B отримані за допомогою k методів, то спочатку обчислюється середньоквадратична похибка, потім значення ранжуються від найменшого до найбільшого. Хоча непараметричні методи мають свої переваги, важливо усвідомлювати, що вони ігнорують частину доступної інформації. Так, критерії міток та рангів не враховують числові значення похибок .

Висновки

Таким чином, дослідження процесу інтелектуального аналізу даних показало, що розширення інструментарію аналізу даних у зв'язку з потужним розвитком технологій, формуванням великих масивів даних створює можливість відслідковувати, оцінювати, моделювати та, зрештою, враховувати ключові економічні та соціальні зміни і тенденції в складних процесах і системах. Важливим кроком, який підвищив ефективність інтелектуального аналізу даних стало включення кроків, що полегшують отримання даних, а також їхню оцінювання та візуалізації моделей.

Описові та прогнозні моделі, створені в процесі інтелектуального аналізу, можуть і повинні використовуватися разом. Логічна послідовність застосування моделі, яка покращить результати спрямованого моделювання, вбачається, насамперед, у пошуку закономірностей у даних за допомогою дескриптивних моделей, а вже на основі отриманих ідей спрямованого моделювання при створенні прогнозних моделей складних соціально-економічних процесів і систем.

На сучасному етапі розвитку технологій машинне навчання широко використовується в інтелектуальному аналізі даних для винайдення складних моделей та алгоритмів, які слугують для створення описових та прогнозних моделей складних соціально-економічних систем та процесів. Машинне навчання дає комп'ютерам можливість "вчитися", розпізнавати складні закономірності та приймати інтелектуальні рішення без явного програмування на основі великих вибірок даних. Ці можливості є основним застосуванням методів глибокого навчання, призначених для обробки даних, представлених у вигляді багатовимірних масивів, і дозволяють створювати моделі складних соціально-економічних процесів і враховувати можливі зміни для проектування та управління розвитком складних систем. Тобто, використання вищезазначених інструментів дозволяє успішно та ефективно виконувати завдання інтелектуального аналізу даних складних соціально-економічних процесів та систем.

Тому цифрові інструменти стають актуальними для підтримки ефективної конкурентоспроможності, допомагають моделювати складні соціально-економічні процеси та системи, ефективно аналізувати та використовувати існуючі великі масиви даних для оперативного управління людськими ресурсами та стратегічного планування складних соціально-економічних процесів і систем.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Лопатко О. Нейронні мережі як засіб прогнозування значення температури за перехідним процесом // Вимірюв. техніка та метрологія : міжвід. наук.–техн. зб. – 2016. – Вип. 77.
2. Wooldridge J. M. Introductory econometrics: a modern approach / J. M. Wooldridge. – 4–th edition. – Mason, OH : Cencage Learning, 2009. – 865 p
3. Кучанський О. Ю. Інформаційна система підтримки прийняття рішень у діяльності фінансових установ на основі трендових моделей : автореф. дис. ... канд. техн. наук : 05.13.06. – Київ, 2014

Мазуренко Владислав Володимирович **–** студент групи 1КІ-22м, факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, Вінниця, e-mail: mazurenkovlad226@gmail.com

Добровольська Наталя Вікторівна **–** доцент, кафедра обчислювальної техніки, Вінницький національний технічний університет, Вінниця