РОЗДІЛ 9
МАТЕМАТИЧНІ МЕТОДИ, МОДЕЛІ ТА ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ В ЕКОНОМІЦІ

УДК 330.4

DOI: https://doi.org/10.32782/2304-0920/3-82-21

Білоскурський Р. Р.
Верстяк А. В.
Чернівецький національний університет імені Юрія Федьковича

ОСОБЛИВОСТІ ФОРМУВАННЯ ІНФОРМАЦІЙНОГО ЗАБЕЗПЕЧЕННЯ АЛГОРИТМІВ МАШИННОГО НАВЧАННЯ В ЗАДАЧАХ З ЕКОНОМІКИ

У статті розглянуто основні підходи та специфіку формування інформаційного забезпечення моделей машинного навчання з учителем в економічних задачах. Досліджено головні етапи підготовки даних для моделювання. Встановлено та класифіковано джерела інформаційного забезпечення: неструктурована зовнішня інформація, неструктурована внутрішня інформація, структуровані датасети з розміченими даними машиночитного формату. Особливу увагу приділено порталам відкритих даних. Запропоновано підходи до роботи із сирою інформацією та очищення вхідної інформації. Детально проаналізовано класифікацію ознак наборів даних алгоритмів машинного навчання та особливості роботи з ними. Розглянуто методи попередньої обробки інформації: нормалізацію, відбір ознак, крос-валідацію, візуалізацію даних.
Ключові слова: цифрова економіка, машинне навчання, інформаційне забезпечення, датасет, аналіз даних, ознака.

Постановка проблеми. Упровадження в процеси цифрового управління технологій блокчейну, математичного та модельного інструментарію Business Intelligence, Data Mining, Data Science, Machine Learning, Artificial Intelligence зумовлюють необхідність суттєвої зміни теоретичної бази та методології досліджень в економіці, зокрема щодо трактування закономірностей розвитку, процесів цілепокладання та обґрунтування рішень. За концепцією Big Data та новітньої аналітичної парадигми цифрової економіки змодельовані закономірності мають первинну форму регулярностей у даних та видобуваються з величезних обсягів переважно неструктурованої інформації, наприклад результатів транзакцій, дослідження профілів і траєкторій мережевої активності учасників ринків [1, с. 334]. Машинне навчання (Machine Learning) – одна з фундаментальних технологій цифрової економіки та Четвертої промислової революції, що базується на побудові алгоритмів, які здатні самонавчатися. Розрізняють три основних способи машинного навчання: навчання з учителем, навчання без учителя та навчання з підкріпленням. Навчання з учителем має найбільше застосування на практиці, але потребує добре підготовлених та розмічених даних.

Аналіз останніх досліджень і публікацій. Застосування методів машинного навчання та інтелектуального аналізу даних в економічних системах здійснювали такі вітчизняні та зарубіжні науковці, як І.В. Мірошниченко, К.Г. Івлієва [2], Н.В. Дунас, М.С. Білокриницька [3], які досліджували оцінювання кредитного ризику банківських установ методами машинного навчання; Susan Athey, Guido W. Imbens [4], Susan Athey [5], Р.В. Шамин [6] аналізували особливості використання алгоритмів машинного навчання в економіці; A. Kamilaris, F.X. Prenafeta-Boldu [7], X. Pham, M. Stack [8] досліджували аналіз даних в агропромисловому секторі.

Виділення не вирішених раніше частин загальної проблеми. Попри значний обсяг досліджень штучного інтелекту, машинного навчання та аналізу даних, що з'явилися у вітчизняних та закордонних наукових виданнях, вивчення специфіки їх застосування для вирішення проблем національної економіки є актуальним завданням. Зокрема, якщо алгоритми машинного навчання загалом незмінні для будь-якого об'єкта дослідження, то інформаційне забезпечення моделювання має свою специфіку. При цьому інформаційне забезпечення в задачах машинного навчання з учителем відіграє вирішальне значення, оскільки отримані закономірності на відміну від аналітичних підходів базуються на аналізі даних.

Мета статті. Головною метою цієї роботи є дослідження та обґрунтування методики формування релевантного інформаційного забезпечення систем машинного навчання у моделях економічних явищ та процесів.

Виклад основного матеріалу. Інформаційне забезпечення разом із програмним, технічним, лінгвістичним, організаційним та документальним забезпеченням є базовим компонентом довільної системи обробки економічної інформації. У науковій літературі триває дискурс стосовно визначення поняття «інформаційне забезпечення», оскільки різні автори використовують різні підходи та точки зору стосовно дефініції «інформація»: філософський (універсальна субстанція, що пронизує усі сфери людської діяльності, слугує провідником знань та думок, інструментом спілкування, взаєморозуміння та співробітництва, утвердження стереотипів мислення та поведінки), кібернетичний (комунікація та зв'язок, у процесі якого усувається інформаційна ентропія) і навіть інтуїтивний (набір відомостей про навколишній світ). У широкому розумінні інформаційне забезпечення – це система організації збору, форматування, обробки та використання даних в інформаційних системах із використанням від-

повідного програмного забезпечення. У системах машинного навчання під інформаційним забезпеченням розуміють розмічені, добре структуровані набори табличних даних, які називаються датасетами. Рядки датасету називаються об'єктами, а стовпці – ознаками. Як правило, файли датасетів алгоритмів машинного навчання зберігаються у форматах CSV (comma-separated values), Parquet, Avro, Protobuf, що зручні для обробки відповідними мовами програмування (Python, R, C++, JavaScript, Java).

Існує два основних підходи до збору даних задач машинного навчання економічних систем: формування власної інформаційної бази з неструктурованої інформації та використання існуючих датасетів. При цьому важливо чітко розуміти бізнес-задачу, для якої буде реалізовуватися модель машинного навчання. Формування власної інформаційної бази може відбуватися як на основі офіційної статистики, так і з використанням різноманітних електронних датчиків та сенсорів технологічних процесів, систем анкетування та опитування, автоматичної онлайн-реєстрації даних у процесі замовлення та покупок в Інтернет-магазинах, парсингу вебсайтів, у тому числі і соціальних мереж. Парсинг соціальних мереж дає змогу визначити тренди та тенденції на ринку, провести сегментацію цільової аудиторії, відслідковувати настрій споживачів та їхню лояльність до бренду. Також добре зарекомендували себе сервіси на зразок Google Trends [9], який показує, як часто певний термін шукають по відношенню до загального обсягу пошукових запитів у різних регіонах світу і на різних мовах та Google Analytics [10] для отримання детальної аналітики аудиторії сайтів, стану індексування та оптимізації видимості у пошукових сервісах. Зауважимо, що аналіз більшої кількості джерел збільшує шанси на виявлення прихованих закономірностей та кореляції в даних.

Важливим джерелом отримання інформаційного забезпечення є портали відкритих даних. Так, в Україні, відповідно до Закону «Про доступ до публічної інформації», було розроблено Єдиний державний вебпортал відкритих даних з електронною адресою data.gov.ua [11]. Станом на травень 2020 р. на сайті було представлено 30 960 наборів даних. Структура сайту дає змогу здійснювати пошук наборів даних за такими групами: будівництво, держава, екологія, економіка, земля, молодь і спорт, освіта та культура, охорона здоров'я, податки, сільське господарство, соціальний захист, стандарти, транспорт, фінанси, юстиція. Також локальні портали даних упроваджують багато українських міст. Зокрема, лідерами є Харків, Львів, Дніпро, Дрогобич. На міжнародному рівні відкриті дані доступні на багатьох ресурсах. Зокрема, Європейський портал даних [12] містить понад 1 млн найрізноманітніших датасетів із 35 країн Європи, що згруповані у 85 каталогів; сайт Світового банку [13] – понад 20 тис наборів даних, що охоплюють економічні показники та індикатори; сайт Міжнародного валютного фонду [14] містить датасети про міжнародні фінанси, валютні резерви, показники боргу, інвестицій, інфляції, ціни на сировинних ринках; сайт газети Financial Times [15] зберігає набори даних у форматі часових рядів про світові фінансові ринки.

Разом із тим слід зауважити, що значна частина інформації з порталів відкритих даних не може без попередньої обробки використовуватися в алгоритмах машинного навчання. Це пов'язано з незручними форматами представлення даних (pdf, doc, rtf, jpg), які не дають змоги здійснювати обробку файлів.

Окрім зручного формату, набори даних повинні бути очищеними. Чистими є дані, що мають такі характеристики: використання уніфікованого формату для різних типів даних (дати, числа, валюти тощо), застосування єдиного типу даних у межах однієї ознаки, використання змінної *NaN* для значень, які пропущені, відсутність дублікатів значень об'єктів, відсутність порожніх рядків та стовпців, відсутність помилок та описок. Для очистки даних використовують як вбудовані методи систем керування базами даних, так і власні розробки аналітиків.

Формалізовано ознакою є відображення $f: X \to D_f$, де D_f – множина допустимих значень ознаки. Якщо задані ознаки $f_1, f_2, \ldots, f_n$, то вектор $x = \left(f_1(x), f_2(x), \ldots, f_n(x)\right)$ називається вектором ознак об'єкта $x \in X$. Під час попередньої обробки ознак необхідно звернути увагу на такі можливі проблеми:

– недостатня розмірність вибірки. Уведення в модель кожної нової ознаки вимагає збільшення розмірності вибірки на порядок, щоб уникнути проблеми недонавчання;

– розріджені набори даних, що характеризуються великою кількістю нульових значень;

– пошук та виявлення «аномалій» – ситуацій, коли ознака набуває абсолютно нетипового значення, що може сигналізувати про інший механізм її формування, наприклад шахрайські

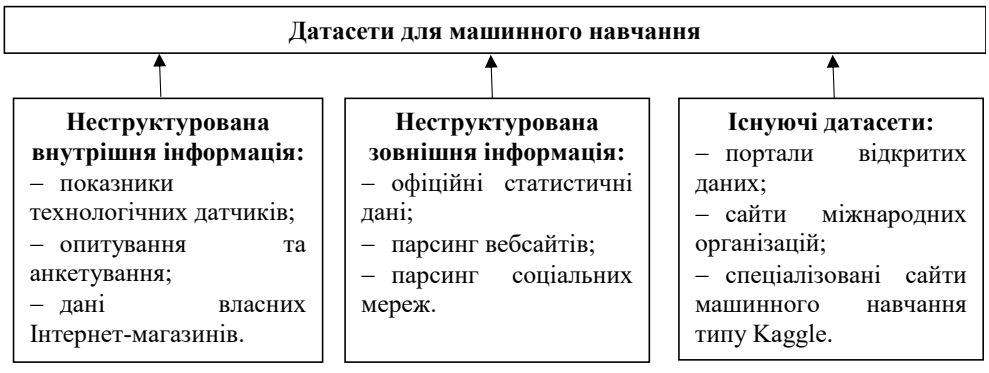

Рис. 1. Джерела інформаційного забезпечення алгоритмів машинного навчання

Джерело: авторська розробка

транзакції з банківськими картками, нетипові значення показників електронних датчиків, різка зміна поведінки у соціальних мережах;

– дублювання ознак, якщо набір даних формувався в результаті об'єднання кількох інформаційних джерел;

– незбалансованість даних, що трапляється, коли певне значення деякої ознаки зустрічається набагато частіше, ніж інші.

Розглянемо основні класи ознак у машинному навчанні, що визначаються виглядом множини допустимих значень D_f:

– бінарні ознаки $D_f = \{0,1\}$. Прикладами бінарних ознак є повернення чи неповернення кредитних коштів позичальником банку, здійснення чи нездійснення покупки клієнтом магазину, наявність чи відсутність знижки. Як правило, ознака приймає значення $\{1\}$ у позитивному варіанті та $\{0\}$ – у негативному;

– дійсні ознаки $D_f = R$. Приклади дійсних ознак: вартість товару, площа будинку, сума прибутку (збитку), кількість відвідувань вебсайту. Якщо для певного об'єкта значення пропущене (NaN), його рекомендовано замінити нулем або середнім арифметичним значенням ознаки за всіма полями;

– порядкові ознаки D_f – скінченна упорядкована множина. Рівні вищої освіти (початковий, бакалаврський, магістерський, докторський), розмір підприємства, категорії водійських прав є прикладами порядкових ознак;

– категоріальні (номінальні) ознаки D_f – неупорядкована множина. Прикладами таких ознак є поштові адреси, кольори, категорії товарів, жанр кінофільмів;

– текстові ознаки D_f – множина символів або слів. Для роботи з текстовими ознаками попередньо необхідно виконати операції токенізації (розбиття вихідного тексту на окремі абзаци, речення, слова), стеммізації (виділення коренів слів шляхом видалення префіксів, суфіксів, закінчень), лемматизації (приведення до канонічної форми слова), видалення слів, які не мають змістовного навантаження (сполучники, прийменники, частки). Зокрема, робота з текстовими ознаками актуальна під час аналізу коментарів користувачів соціальних мереж та відгуків клієнтів сайтів електронної комерції;

– множиннозначні ознаки D_f – множина, елементами якої є інші множини.

Важливим етапом підготовки інформаційного забезпечення є нормалізація даних, що полягає у приведенні дійсних ознак до деякого заданого вузького діапазону значень (як правило [-1, 1] або [0, 1]). Процедура нормалізації даних зумовлена тим, що значення ознак можуть змінюватися дуже сильно і відрізнятися на декілька порядків. Наприклад, вік клієнта банку та його річний дохід. Маючи різну природу походження, дані сильно відрізняються за абсолютними значеннями, що може призвести до неправильної роботи певних методів машинного навчання.

Для того щоб уникнути явища мультиколінеарності, необхідно відкинути «зайві» ознаки. Також це дасть змогу зменшити розмірність моделі, що суттєво впливає на машинний час моделювання. Виділяють такі методи відбору ознак:

– методи фільтрації, які використовують теорію ймовірності та статистику для встановлення фільтрів і оцінюють степінь кореляції кожної ознаки із цільовою змінною;

– обгорткові методи, що ґрунтуються на переборі усіх комбінацій ознак, до яких застосовують машинний алгоритм та кожного разу вимірюють якість його роботи. Такий підхід дає змогу добре підібрати ознаки, але вимагає значних апаратних ресурсів та часу виконання;

– методи регуляризації, які передбачають додавання додаткових обмежень, що штрафують модель за складність. Основними видами регуляризації є lasso (L_1-регуляризація) та ridge (L_2-регуляризація).

Методи регуляризації дають змогу зменшити проблему перенавчання моделей, яка полягає у тому, що алгоритм демонструє дуже добрі результати під час навчання на відомих розмічених даних та погані – на нових даних. Проте, загалом зменшуючи проблему перенавчання, регуляризація не дає відповіді на запитання про якість роботи моделі та відсоток помилок на нових наборах даних.

Найпростіший спосіб оцінки якості алгоритму машинного навчання полягає у розбитті датасету на дві частини. Перша з них – тренувальна, використовується для навчання алгоритму, а друга – тестова, для перевірки якості моделі, шляхом оцінки метрик якості. Такий підхід називається методом відкладеної вибірки. Як правило, тренувальна вибірка становить близько 70–80% даних, а тестова, відповідно, – 30–20%. При цьому важливо, щоб вибірка була добре «перемішаною» й однакові чи схожі об'єкти рівномірно розподілялися у тренувальній та тестовій вибірках. Винятком є часові ряди у задачах аналізу фінансових ринків, де тренувальна вибірка завжди складається з більш давніх даних, а тестова – з нових. Інакше ми будемо тренувати модель на даних «із майбутнього».

Метод відкладеної вибірки можна вдосконалити, розбиваючи датасет на k (на практиці k вибирають із множини $\{3,4,\ldots,10\}$ залежно від розміру вибірки) однакових за розмірами блоків. Далі кожен із блоків використовують як тестовий, а інші – як тренувальну вибірку. Таким чином, знаходячи середнє арифметичне k варіантів якості алгоритму, отримуємо кінцевий результат. Такий підхід називається крос-валідацією.

Заключним етапом підготовки інформаційного забезпечення може бути візуалізація даних, яка забезпечує графічне відображення даних у вигляді графіків, діаграм, дашбордів.

Висновки і пропозиції. Необхідність фундаментального підходу до формування інформаційного забезпечення алгоритмів машинного моделювання зумовлена критичною важливістю даних для коректної роботи моделей. Без попередньої підготовки інформації, очистки та структуризації даних, інженерії ознак неможливе застосування методів штучного інтелекту, а саме машинного навчання з учителем. Запропоновані підходи до роботи із сирою інформацією дають змогу здійснювати побудову якісних датасетів для практичного використання в економіці.

Список використаних джерел:
1. Вітлінський В.В., Катуніна О.С. Методологічні аспекти моделювання розвитку та життєздатності систем і контрагентів цифрової економіки. *Проблеми економіки*. 2018. № 1. С. 333–341.
2. Мірошниченко І.В., Івлієва К.Г. Оцінювання кредитного ризику методами машинного навчання. *Ефективна економіка*. 2019. № 12. URL : http://www.economy.nayka.com.ua/?op=1&z=7513 (дата звернення: 01.06.2020).
3. Дунас Н.В., Білокриницька М.С. Впровадження системи кредитного скорингу українськими банками для споживчого кредитування. *Приазовський економічний вісник*. 2019. № 5(16). С. 263–269.
4. Susan Athey, Guido W. Imbens. Machine Learning Methods That Economists Should Know About. *The Annual Review of Economics*. 2019. № 11. P. 685–725.
5. Susan Athey. The Impact of Machine Learning on Economics. A chapter in The Economics of Artificial Intelligence : An Agenda, 2018. P. 507–547.
6. Шамин Р.В. Машинное обучение в задачах экономики. Москва : Грин Принт, 2019. 140 с.
7. Kamilaris A., Prenafeta-Boldu F.X. Deep learning in agriculture: A survey. *Comput. Electron. Agric.* 2018. № 147. P. 70–90.
8. Pham X., Stack M. How data analytics is transforming agriculture. *Bus. Horiz.* 2018. № 61. P. 125–133.
9. Google Trends. URL : https://trends.google.com (дата звернення: 01.06.2020).
10. Google Analytics. URL : https://analytics.google.com (дата звернення: 01.06.2020).
11. Єдиний державний вебпортал відкритих даних. URL : https://data.gov.ua (дата звернення: 01.06.2020).
12. Європейський портал даних. URL : https://www.europeandataportal.eu (дата звернення: 01.06.2020).
13. Сайт відкритих даних Світового банку. URL : https://data.worldbank.org (дата звернення: 01.06.2020).
14. Сайт даних Міжнародного валютного фонду. URL : https://www.imf.org/en/Data (дата звернення: 01.06.2020).
15. Сайт ринкових даних Financial Times. URL : https://markets.ft.com/data (дата звернення: 01.06.2020).

References:
1. Vitlinskyi V.V., Katunina O.S. (2018). Metodolohichni aspekty modeliuvannia rozvytku ta zhyttiezdatnosti system i kontrahentiv tsyfrovoi ekonomiky. [Methodological aspects of modeling the development and viability of systems and counterparties of the digital economy]. *Problemy ekonomiky*, no. 1. pp. 333–341.
2. Miroshnychenko I.V., Ivliieva K.H. (2019). Otsiniuvannia kredytnoho ryzyku metodamy mashynnoho navchannia [Credit risk assessment by machine learning methods]. *Efektyvna ekonomika*, no. 12. URL: http://www.economy.nayka.com.ua/?op=1&z=7513 (accessed 1 June 2020).
3. Dunas N.V., Bilokrynytska M.S. (2019) Vprovadzhennia systemy kredytnoho skorynhu ukrainskymy bankamy dlia spozhyvchoho kredytuvannia [Introduction of a credit scoring system by Ukrainian banks for consumer lending]. *Pryazovskyi ekonomichnyi visnyk*, no. 5(16). pp. 263-269.
4. Susan Athey, Guido W. Imbens. (2019) Machine Learning Methods That Economists Should Know About. *The Annual Review of Economics*, no. 11. pp. 685–725.
5. Susan Athey. (2018) The Impact of Machine Learning on Economics. A chapter in The Economics of Artificial Intelligence: An Agenda, pp. 507-547.
6. Shamyn R.V. (2019) Mashynnoe obuchenye v zadachakh ekonomyky [Machine Learning in Economics]. «Hryn Prynt», 140 p.
7. Kamilaris, A.; Prenafeta-Boldu, F.X. (2018) Deep learning in agriculture: A survey. *Comput. Electron. Agric.* no. 147, pp. 70-90.
8. Pham, X.; Stack, M. (2018) How data analytics is transforming agriculture. *Bus. Horiz.* no 61, pp. 125–133.
9. Google Trends. URL: https://trends.google.com (accessed 1 June 2020).
10. Google Analytics. URL: https://analytics.google.com (accessed 1 June 2020).
11. Iedynyi derzhavnyi vebportal vidkrytykh danykh [The only state open data web portal]. URL: https://data.gov.ua (accessed 1 June 2020).
12. European Data Portal. URL: https://www.europeandataportal.eu (accessed 1 June 2020).
13. World Bank Open Data. URL: https://data.worldbank.org (accessed 1 June 2020).
14. International Monetary Fund Data. URL: https://www.imf.org/en/Data (accessed 1 June 2020).
15. Markets Data Financial Times. URL: https://markets.ft.com/data (accessed 1 June 2020).

Белоскурский Р. Р.
Верстяк А. В.
Черновицкий национальный университет имени Юрия Федьковича

ОСОБЕННОСТИ ФОРМИРОВАНИЯ ИНФОРМАЦИОННОГО ОБЕСПЕЧЕНИЯ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ В ЗАДАЧАХ ПО ЭКОНОМИКЕ

Резюме
В статье рассмотрены основные подходы и специфика формирования информационного обеспечения моделей машинного обучения с учителем в экономических задачах. Исследованы основные этапы подготовки данных для моделирования. Установлены и классифицированы источники информационного обеспечения: неструктурированная внешняя информация, структурирована внутренняя информация, структурированные датасеты с размеченными данными в машиночитаемом формате. Особое внимание уделено порталам открытых данных. Предложены подходы к работе с сырой информацией и очистке входящей информации. Детально проанализирована классификация признаков наборов данных алгоритмов машинного обучения и особенности работы с ними. Рассмотрены методы предварительной обработки информации: нормализация, отбор признаков, кросс-валидация, визуализация данных.
Ключевые слова: цифровая экономика, машинное обучение, информационное обеспечение, датасет, анализ данных, признак.

Biloskurskyy Ruslan
Verstiak Andrii
Yuriy Fedkovych Chernivtsi National University

SPECIAL ASPECTS OF CREATION OF INFORMATION SUPPORT OF MACHINE LEARNING ALGORITHMS IN ECONOMIC PROBLEMS

Summary
Information support in supervised machine learning tasks is important because the obtained patterns are based on data analysis in contrast to analytical approaches. The main approaches and specifics of the creation of information support of supervised machine learning models in economic problems are considered in the article. The main stages of data preparation for modeling are investigated. Sources of information support are established and classified. There are unstructured external information, unstructured internal information, structured datasets with marked data. The creation of the information base can take place on the basis of official statistics, with the use of various electronic sensors and sensors of technological processes, questionnaire and survey systems, automatic online registration of data in the process of ordering and shopping in online stores, parsing websites, including social networks. Special attention is paid to open data portals. Ukrainian, European and World open data portals were analyzed. Approaches to work with raw information and problems of clearing input information are offered. Pure data has certain characteristics of quality. For pure data features include a uniform format for different data types, single data type within a single attribute, the NaN variable for missing values, no duplicate object values, no empty rows and columns, no errors and omissions. For data cleaning, both built-in methods of database management systems and analysts' own developments are used. The classification of features of datasets of machine learning algorithms are analyzed in detail. When pre-processing the features, it is necessary to pay attention to the following possible problems: insufficient sample size, sparse datasets, characterized by a large number of zero values, search and detection of anomalies, data imbalance. The tools of reducing the dimensionality of machine learning models are investigated. Normalization, selection of features, cross-validation, data visualization methods of preprocessing of information are considered in the article.
Keywords: digital economy, machine learning, information support, dataset, data analysis, feature.