https://doi.org/10.31891/2307-5732-2025-355-24
УДК 004.652

ЄВСІНА НАТАЛЯ
Національний технічний університет «Харківський політехнічний інститут»
http://orcid.org/0000-0003-0214-7383
e-mail: nataliia.yevsina@khpi.edu.ua
ДУДНИК ОЛЕКСІЙ
Національний технічний університет «Харківський політехнічний інститут»
https://orcid.org/0009-0008-6529-0703
e-mail: oleksii.dudnik@khpi.edu.ua
ГАПОН АНАТОЛІЙ
Національний технічний університет «Харківський політехнічний інститут»
https://orcid.org/0000-0002-2582-6154
e-mail:anatolii.hapon@khpi.edu.ua
ЄВСІН ГРИГОРІЙ
Національний технічний університет «Харківський політехнічний інститут»
https://orcid.org/0009-0009-7771-796X
e-mail: hryhorii.yevsin@cit.khpi.edu.ua

АНАЛІЗ МЕТОДІВ ОПРАЦЮВАННЯ ВЕЛИКИХ ДАНИХ В ІНТЕЛЕКТУАЛЬНИХ СИСТЕМАХ КЕРУВАННЯ

У роботі проведено аналіз методів інтелектуальної обробки великих даних, які створюються та накопичуються в інтелектуальних системах керування. Розвиток інформаційних технологій вимагає швидкого управління великими обсягами неструктурованих даних. З появою дедалі більшої кількості IoT-пристроїв і сенсорів, здатних взаємодіяти з навколишнім середовищем і між собою, виникає цифрове середовище, що містить величезні обсяги різноманітних та змінюваних даних. Традиційні методи штучного інтелекту та машинного навчання, розроблені для детермінованих ситуацій, не підходять для таких умов. Оскільки кожен пристрій у інтелектуальній системі керування потребує великої кількості параметрів, важливо, щоб штучний інтелект міг адаптуватися і навчатися без заздалегідь заданої структури і параметрів.

У статті виконано короткий огляд існуючих підходів і характеристик методів Data Mining. Наведено алгоритм процесу інтелектуального опрацювання даних, який дозволяє порівняти ефективність побудови інтелектуальної системи керування.

Ключові слова: Великі дані, Інтелектуальні системи керування, Data Mining, Штучний інтелект, Методи інтелектуального опрацювання даних

YEVSINA NATALIA
DUDNIK ALEXEY
GAPON ANATOLY
YEVSIN HRYHORII
National Technical University «Kharkiv Polytechnic Institute»

ANALYSIS OF BIG DATA PROCESSING METHODS IN INTELLIGENT CONTROL SYSTEMS

Intelligent data analysis is one of the key areas of artificial intelligence (AI) that uses AI methods and principles to automatically discover hidden patterns and knowledge in large datasets. The main idea of this field is to develop algorithms and models capable of automatically analyzing, classifying, clustering, and processing data in such a way that relationships and knowledge, which are not visible to the naked eye, can be revealed.

The goal of intelligent data processing, like AI in general, is to automate the data analysis process and uncover complex dependencies and knowledge in intelligent control systems. This approach is especially important when the volume of data is enormous, the information is diverse and complex, and complex mathematical and statistical methods are required to discover hidden patterns.

The article analyzes the methods of intelligent processing of big data that are created and accumulated in intelligent control systems. The development of information technologies requires rapid management of large volumes of unstructured data. With the emergence of an increasing number of IoT devices and sensors capable of interacting with the environment and with each other, a digital environment is emerging that contains huge volumes of diverse and changing data. Traditional methods of artificial intelligence and machine learning, developed for deterministic situations, are not suitable for such conditions. Since each device in an intelligent control system requires a large number of parameters, it is important that artificial intelligence can adapt and learn without a predetermined structure and parameters.

The article provides a brief overview of existing approaches and characteristics of Data Mining methods. An algorithm for the process of intelligent data processing is presented, which allows comparing the efficiency of building an intelligent control system.

Keywords: Big data, Intelligent control systems, Data Mining, Artificial intelligence, Methods of intelligent data processing

Стаття надійшла до редакції / Received 07.05.2025
Прийнята до друку / Accepted 06.06.2025

Постановка проблеми у загальному вигляді та її зв'язок із важливими науковими чи практичними завданнями

Інтелектуальне опрацювання даних (Data Mining) базується на концепції виявлення шаблонів, які відображають різноманітні взаємозв'язки в даних. Ці шаблони є закономірностями, властивими певним підмножинам даних і можуть бути перетворені на зрозумілу людині форму. Процес пошуку шаблонів здійснюється методами, які не обмежуються заздалегідь заданими припущеннями про структуру даних або

типи розподілів значень аналізованих параметрів. Однією з ключових особливостей Data Mining є здатність виявляти неочевидні та несподівані закономірності в даних, що дозволяє знаходити «приховані знання».

Інтелектуальний аналіз даних є однією з основних галузей штучного інтелекту (ШІ), що використовує методи та принципи ШІ для автоматичного виявлення прихованих закономірностей і знань у великих масивах даних. Основна ідея цього напрямку полягає в розробці алгоритмів і моделей, здатних автоматично аналізувати, класифікувати, кластеризувати та обробляти дані таким чином, щоб виявляти зв'язки та знання, які неможливо побачити неозброєним оком.

Термін Big Data введений доктором з інформатики університету Берклі Кліффордом Лінчем [1] у 2008 році. Фундаментальним дослідженням у сфері великих даних присвячені роботи В. Майер-Шенбергера та К. Кук'єра , Ж.-П. Дейкса [2].

Мета інтелектуального опрацювання даних, як і ШІ загалом, полягає в автоматизації процесу аналізу даних і виявленні складних залежностей і знань в інтелектуальних системах керування. Цей підхід особливо важливий, коли обсяг даних величезний, інформація різноманітна та складна, а для знаходження прихованих закономірностей необхідно застосовувати складні математичні і статистичні методи.

Аналіз досліджень та публікацій

Для обробки великих даних існує дві категорії систем: пакетна обробка великих даних і технологія потокової обробки [3]. Широко використовувані технології великих даних були значною мірою прийняті для стимулювання економічного зростання багатьох галузей, включаючи Інтернет-індустрію, а також традиційні галузі виробництва. Наразі типова технічна архітектура обробки великих даних включає Hadoop [4] і похідні від нього системи. Технічна архітектура Hadoop, в основному підтримується Yahoo! та Facebook, реалізує та оптимізує структуру MapReduce [5]. З моменту випуску в 2006 році технічна архітектура Hadoop перетворилася на динамічну екосистему, що містить понад шістдесят компонентів. Крім того, екосистема значно сприяла дослідженням розподілених обчислень [6], хмарних обчислень [7,8] і пов'язаних з ними робіт [9].

У напрямку оцінки продуктивності платформ обробки великих даних дослідники провели велику роботу та розробили різноманітні метрики [10]. Для кількісної оцінки операційної здатності великомасштабних систем аналітичної обробки було розроблено показник, названий базовими операціями за секунду, а в [11] він представлений для оцінки центрів обробки даних. З огляду на різницю обчислювальної потужності та різноманітність навантаження вузлів кластера, запропоновано алгоритм планування з використанням розподілу ресурсів.

Якісне прогнозування – це вид фінансового прогнозування, який спирається на інтуїцію та суб'єктивність. Експерти використовують історичні та сучасні факти, щоб передбачити майбутні закономірності розвитку та правила фінансових операцій, спираючись на знання предмета [12]. Він приймає розумні припущення проблеми як передумову, використовує ймовірнісні та статистичні методи для проведення теоретичного виведення та отримання остаточної моделі прогнозування. Прогнозування часових рядів, регресійне прогнозування, ймовірнісне прогнозування та комбіноване прогнозування є одними з найпоширеніших методів у цій категорії [13].

Коли справа доходить до інтелектуального опрацювання даних, це не окремий предмет, а скоріше сукупність пов'язаних ідей і методів із широкого спектру галузей [14,15].

Формулювання цілей статті

Метою роботи є визначення алгоритму процесу інтелектуального опрацювання даних, який дозволяє порівняти ефективність побудови інтелектуальної системи керування.

Виклад основного матеріалу

Основною рисою Data Mining є поєднання широкого спектра математичних інструментів (від традиційного статистичного аналізу до новітніх кібернетичних методів) з останніми досягненнями в галузі інформаційних технологій. У технології Data Mining гармонійно взаємодіють формалізовані методи та неформальний аналіз, тобто кількісне і якісне дослідження даних.

З розвитком і ускладненням інтелектуальних систем особливого значення набувають підходи, які базуються на побудові систем у вигляді шарів або рівнів. Багаторівнева структура — це модель взаємодії, де компоненти або цілі інтелектуальні системи обмінюються знаннями через певне внутрішнє представлення.

Концепція рівнів дозволяє розділяти складні системи на більш прості, логічно завершені частини. При цьому верхній рівень описує функціонування всієї системи загалом, а нижчі — реалізують конкретні деталі, необхідні для роботи вищих рівнів. Тобто, кожен нижній рівень забезпечує можливості, які використовуються рівнем, що знаходиться над ним.

Поділ системи на рівні має кілька переваг: кожен рівень можна розглядати як самостійний і незалежний; можна реалізувати іншу версію одного з рівнів без змін в інших; залежності між рівнями мінімізовані; один рівень може служити базою для кількох інших.

Однак такий підхід має і недоліки: хоча рівні добре інкапсулюють функції, зміни в одному можуть спричинити ланцюгові зміни в інших; надмірна кількість рівнів може погіршити продуктивність через потребу трансформувати дані між рівнями.

Найбільша складність при розробці багаторівневих систем — це чітко визначити зміст і межі відповідальності кожного окремого рівня.

Створення структури інтелектуальної системи насамперед пов’язане з формуванням моделі цієї системи. У такій моделі мають бути передбачені компоненти, притаманні звичайним інформаційним системам, так і механізми обробки знань, які реалізуються за допомогою інтелектуальних технологій. На відміну від традиційних систем, в інтелектуальних додаються нові елементи — процеси перетворення знань або їх управління, що базуються на застосуванні засобів штучного інтелекту, таких як експертні системи, бази знань, системи підтримки прийняття рішень, асоціативна пам’ять, нечітка логіка тощо.

У зв’язку з цим сучасна складність і багатоваріантність методологій розробки програмного забезпечення часто не дозволяє ефективно вирішувати нові, комплексні завдання, які складаються з численних підзадач і потребують різних підходів. Сучасні вимоги висувають потребу в новому типі програмних рішень — адаптивних, надійних, гнучких, розподілених та відкритих систем.

Задачі Data Mining можна поділити на дескриптивні та прогнозуючі в залежності від використовуваних моделей. Описові задачі (descriptive) дозволяють аналітику виявити шаблони, які описують дані і можуть бути інтерпретовані. Ці задачі фокусуються на загальній концепції аналізованих даних, виявляючи їх інформативні, підсумкові та характерні ознаки.

Прогнозуючі (predictive) задачі, в свою чергу, базуються на аналізі даних для створення моделей, прогнозуванні тенденцій або властивостей нових чи невідомих даних. Однією з важливих особливостей Data Mining є нетривіальність виявлених шаблонів. Це означає, що виявлені шаблони повинні відображати неочевидні та несподівані закономірності в даних, що складають так звані приховані знання (hidden knowledge).

Не існує єдиної думки щодо того, які саме завдання належать до Data Mining, але більшість авторитетних джерел виділяє такі основні задачі, як класифікація, кластеризація, прогнозування, асоціація, візуалізація та підведення підсумків (таблиця 1).

Таблиця 1

Характеристика методів Data Mining

Назва	Характеристика
Класифікація(Classification)	Це одна з найбільш базових і широко використовуваних задач у Data Mining. Для її вирішення застосовуються різні методи, такі як метод найближчого сусіда (Nearest Neighbor), алгоритм к-ближчих сусідів (k-Nearest Neighbor) і нейронні мережі (neural networks).
Кластеризація (Clustering)	Кластеризація є розвитком ідеї класифікації, але вона складніша. Основна її відмінність полягає в тому, що спочатку класи об'єктів не визначені, і задача полягає в тому, щоб виявити природні групи чи кластери серед даних.
Асоціація (Associations)	Задача асоціації полягає у пошуку закономірностей між різними подіями чи елементами в наборі даних. Це дозволяє виявити залежності між елементами, які часто трапляються разом, і побудувати асоціативні правила.
Послідовність (Sequence)	Ця задача зосереджена на виявленні часових закономірностей, тобто вона дозволяє знаходити залежності між транзакціями або подіями, що відбуваються в певній послідовності або часі.
Прогнозування (Forecasting).	Прогнозування є задачею, в якій на основі аналізу наявних даних визначаються пропущені або майбутні значення певних числових показників. Це важливий інструмент для оцінки майбутніх тенденцій чи подій.
Візуалізація (Visualization, Graph Mining)	Процес візуалізації передбачає створення графічних образів даних для кращого розуміння їх структури та зв'язків. Цей метод дозволяє виявляти закономірності в даних через візуальні елементи, такі як графіки чи діаграми.
Підведення підсумків (Summarization)	Задача підведення підсумків спрямована на створення короткого опису або узагальнення даних для певних груп об'єктів. Це допомагає виявити загальні риси або характеристики груп об'єктів, що дозволяє краще зрозуміти набір даних в цілому.

Інтелектуальне опрацювання даних є великим та комплексним процесом, який включає всі етапи — від постановки запитань щодо даних та розробки моделей для їх аналізу до впровадження цих моделей у реальне робоче середовище. Цей процес можна розглядати як послідовність із шести основних кроків.

Перший крок полягає в аналізі вимог, визначенні проблемної області, метрик для оцінки моделі, а також у формулюванні завдань для проекту інтелектуального аналізу даних. Другим кроком є об’єднання та очищення даних. Дані можуть зберігатися у різних форматах або містити помилки, такі як некоректні записи. Очищення даних полягає не тільки в видаленні неприпустимих елементів, а й у пошуку прихованих залежностей, визначенні точних джерел даних та підборі найбільш підходящих ознак для аналізу. Важливо також виявити, які фактори найбільше впливають на результати моделі, або які на вигляд

незалежні, але насправді мають сильний взаємозв'язок, що може непередбачувано вплинути на кінцеві результати. Тому перед розробкою моделей слід визначити й усунути такі проблеми.

Третій крок полягає в перегляді підготовлених даних. Для ефективної розробки моделей необхідно ретельно вивчити дані. Методи дослідження включають обчислення мінімальних і максимальних значень, стандартного відхилення та аналіз розподілу даних. Наприклад, розглядаючи максимальні, мінімальні та середні значення, можна зробити висновок, що вибірка є непредставницькою, і для досягнення кращих результатів потрібні більш збалансовані дані або коригування припущень щодо очікуваних результатів. Стандартне відхилення та інші характеристики розподілу допомагають оцінити стабільність та точність моделі. Високе стандартне відхилення може свідчити, що додавання нових даних покращить модель.

Четвертий крок включає побудову моделей інтелектуального аналізу даних. Отримані під час перегляду даних висновки допомагають визначити, які моделі найкраще підходять для вирішення поставлених завдань. До цього моменту модель є лише контейнером, що визначає параметри для обробки даних. Обробка, або навчання моделі, включає застосування математичного алгоритму до даних з метою виявлення закономірностей. Параметри алгоритму налаштовуються відповідно до потреб, а фільтри допомагають вибрати тільки певну підмножину даних, що дає різні результати.

П’ятий крок — це перевірка ефективності побудованих моделей. Необхідно оцінити, наскільки добре кожна модель працює для даного завдання. Під час розробки можуть бути створені кілька моделей з різними налаштуваннями, і їх перевіряють, щоб обрати найбільш ефективну. Якщо жодна модель не дає бажаних результатів, може знадобитися повернення до попередніх етапів для коригування завдання або переробки аналізу даних.

Останнім етапом є розгортання найбільш ефективних моделей у реальному робочому середовищі. Важливо, що кожен крок не завжди веде до наступного, тому процес створення інтелектуальної системи для аналізу даних є динамічним та ітеративним. Після перегляду даних може з’ясуватися, що для побудови моделей необхідно шукати додаткові дані, або може виникнути потреба в оновленні існуючих моделей новими даними.

Висновки з даного дослідження і перспективи подальших розвідок у даному напрямі

Рішення складних та різноманітних завдань вимагає використання досягнень у сфері штучного інтелекту та розробок у методології і технології створення інтелектуальних систем керування.

Методи, що використовуються в різних інтелектуальних системах, не є універсальними. Переваги одних методів компенсують недоліки інших через їх взаємодію, що дає нові інтегративні властивості. Таким чином, інтеграція різних методів дозволяє подолати обмеження кожного окремого методу, створюючи кілька можливостей для обробки інформації в межах однієї архітектури.

Більшість аналітичних методів, застосовуваних у Data Mining, є добре відомими математичними алгоритмами, але їх нове застосування пов'язане з можливістю вирішення конкретних завдань завдяки новим технічним і програмним можливостям.

Література

1. Lynch, C. A. (2008). Big data: How do your data grow. *Nature, 455*(7209), 28. https://doi.org/10.1038/455028a

2. Special Issue: Big data in communication research. (2014). *Journal of Communication, 64*(2), 193–360, E1–E9. https://doi.org/10.1111/jcom.12090

3. Chen, C. (2017). Real-time processing technology, platform and application of streaming big data. *Big Data, 3*, 1–8. https://doi.org/10.1007/s11432-018-9834-8

4. Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010). The Hadoop distributed file system. In *Proceedings of the IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)* (pp. 1–10). https://doi.org/10.1109/MSST.2010.5496972

5. Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. *Communications of the ACM, 51*(1), 107–113. https://doi.org/10.1145/1327452.1327492

6. Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010). Spark: Cluster computing with working sets. *HotCloud, 10*, 95. https://doi.org/10.1007/978-3-319-14325-5_12

7. Zhang, Q., Cheng, L., & Boutaba, R. (2010). Cloud computing: State-of-the-art and research challenges. *Journal of Internet Services and Applications, 1*(1), 7–18. https://doi.org/10.1007/s13174-010-0007-6

8. Hashem, I. A. T., Yaqoob, I., Anuar, N. B., et al. (2015). The rise of “big data” on cloud computing: Review and open research issues. *Information Systems, 47*, 98–115. https://doi.org/10.1016/j.is.2014.07.006

9. Wu, Q., Ishikawa, F., Zhu, Q., et al. (2017). Deadline-constrained cost optimization approaches for workflow scheduling in clouds. *IEEE Transactions on Parallel and Distributed Systems, 28*(12), 3401–3412. https://doi.org/10.1109/TPDS.2017.2735400

10. Zhao, X., Garg, S., Queiroz, C., et al. (2017). A taxonomy and survey of stream processing systems. In *Software Architecture for Big Data and the Cloud* (pp. 183–206). https://doi.org/10.1016/B978-0-12-805467-3.00011-9

11. Ali, M. (2010). An introduction to Microsoft SQL Server StreamInsight. In *Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research & Application* (p. 66). https://doi.org/10.1145/1823854.1823929

12. Hyde, J. (2010). Data in flight. *Communications of the ACM, 53*(10), 48–52. https://doi.org/10.1145/1629175.1629195

13. Demers, A. J., Gehrke, J., Panda, B., et al. (2007). Cayuga: A general purpose event monitoring system. In *Proceedings of the 3rd Biennial Conference on Innovative Data Systems Research* (Vol. 7, pp. 412–422). https://doi.org/10.1007/s11432-018-9834-8

14. Strohbach, M., Ziekow, H., Gazis, V., et al. (2015). Towards a big data analytics framework for IoT and smart city applications. In *Modeling and Processing for Next-generation Big-data Technologies* (pp. 257–282). https://doi.org/10.1007/978-3-319-09177-8_11

15. Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing, 234*, 11–26. https://doi.org/10.1016/j.neucom.2016.12.038

References

1. Lynch, C. A. (2008). Big data: How do your data grow. Nature, 455(7209), 28. https://doi.org/10.1038/455028a

2. Special Issue: Big data in communication research. (2014). Journal of Communication, 64(2), 193–360, E1–E9. https://doi.org/10.1111/jcom.12090

3. Chen, C. (2017). Real-time processing technology, platform and application of streaming big data. Big Data, 3, 1–8. https://doi.org/10.1007/s11432-018-9834-8

4. Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010). The Hadoop distributed file system. In Proceedings of the IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST) (pp. 1–10). https://doi.org/10.1109/MSST.2010.5496972

5. Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. Communications of the ACM, 51(1), 107–113. https://doi.org/10.1145/1327452.1327492

6. Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010). Spark: Cluster computing with working sets. HotCloud, 10, 95. https://doi.org/10.1007/978-3-319-14325-5_12

7. Zhang, Q., Cheng, L., & Boutaba, R. (2010). Cloud computing: State-of-the-art and research challenges. Journal of Internet Services and Applications, 1(1), 7–18. https://doi.org/10.1007/s13174-010-0007-6

8. Hashem, I. A. T., Yaqoob, I., Anuar, N. B., et al. (2015). The rise of “big data” on cloud computing: Review and open research issues. Information Systems, 47, 98–115. https://doi.org/10.1016/j.is.2014.07.006

9. Wu, Q., Ishikawa, F., Zhu, Q., et al. (2017). Deadline-constrained cost optimization approaches for workflow scheduling in clouds. IEEE Transactions on Parallel and Distributed Systems, 28(12), 3401–3412. https://doi.org/10.1109/TPDS.2017.2735400

10. Zhao, X., Garg, S., Queiroz, C., et al. (2017). A taxonomy and survey of stream processing systems. In Software Architecture for Big Data and the Cloud (pp. 183–206). https://doi.org/10.1016/B978-0-12-805467-3.00011-9

11. Ali, M. (2010). An introduction to Microsoft SQL Server StreamInsight. In Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research & Application (p. 66). https://doi.org/10.1145/1823854.1823929

12. Hyde, J. (2010). Data in flight. Communications of the ACM, 53(10), 48–52. https://doi.org/10.1145/1629175.1629195

13. Demers, A. J., Gehrke, J., Panda, B., et al. (2007). Cayuga: A general purpose event monitoring system. In Proceedings of the 3rd Biennial Conference on Innovative Data Systems Research (Vol. 7, pp. 412–422). https://doi.org/10.1007/s11432-018-9834-8

14. Strohbach, M., Ziekow, H., Gazis, V., et al. (2015). Towards a big data analytics framework for IoT and smart city applications. In Modeling and Processing for Next-generation Big-data Technologies (pp. 257–282). https://doi.org/10.1007/978-3-319-09177-8_11

15. Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. Neurocomputing, 234, 11–26. https://doi.org/10.1016/j.neucom.2016.12.038