АЛГОРИТМІЧНА ТЕОРІЯ ІНФОРМАЦІЇ: ПРОБЛЕМА КОЛМОГОРОВСЬКОЇ СКЛАДНОСТІ В АНАЛІЗІ ДАНИХ

Глотова Анастасія
студентка фізико-математичного факультету
Житомирський державний університет
імені Івана Франка, м. Житомир

Мельник Анна Віталіївна
кандидат педагогічних наук
доцент кафедри комп'ютерних наук та інформаційних технологій
Житомирський державний університет
імені Івана Франка, м. Житомир

У сучасному світі інформаційні технології відіграють ключову роль у розвитку суспільства. Щодня створюються величезні обсяги даних — від соціальних мереж до наукових досліджень і складних фінансових систем. Проте головною проблемою сучасної науки є не лише ефективне зберігання цих даних, а їх глибокий інтелектуальний аналіз та розуміння прихованих структур.

У цьому контексті виникає фундаментальне питання: як об'єктивно визначити, які дані є змістовними (детермінованими), а які — випадковим шумом або зайвою інформацією? Одним із найбільш прогресивних підходів до розв'язання цієї проблеми є алгоритмічна теорія інформації (АТІ). Вона дозволяє оцінювати внутрішню складність інформації незалежно від її семантичного змісту або контексту передачі [3, с. 2].

Центральним поняттям цієї теорії є Колмогоровська складність — міра, що визначає кількість інформації в об'єкті через довжину найкоротшого алгоритму (програми), здатного відтворити цей об'єкт. На основі цієї ідеї було сформульовано принципи, які сьогодні активно

використовуються в сучасних системах машинного навчання та аналізу великих даних [2].

Колмогоровська складність виступає універсальним теоретичним інструментом для оцінки інформації та виявлення закономірностей у даних, проте її практичне впровадження залишається обмеженим через математичну необчислюваність та необхідність використання апроксимаційних (наближених) методів.

Алгоритмічна теорія інформації розглядає інформацію не як ймовірнісну характеристику джерела (як у класичній теорії К. Шеннона), а як результат роботи певного обчислювального процесу. Шеннонівський підхід фокусується на середній очікуваній кількості інформації в повідомленнях від випадкового джерела, тоді як підхід Колмогорова дозволяє оцінювати складність окремого, фіксованого об'єкта.

Колмогоровська складність $K(x)$ об'єкта x визначається як довжина найкоротшої програми p, яка при виконанні на універсальній машині Тюрінга U видає на вихід рядок x:

$$K(x) = min\{|p| : U(p) = x\}$$

Це означає, що чим коротшим є опис об'єкта, тим менш складним він є з алгоритмічної точки зору. Наприклад, послідовність «101010101010» можна компактно описати як «повторити '10' шість разів». Її складність низька. Натомість абсолютно випадкова послідовність бітів не має внутрішніх закономірностей, тому її найкоротший опис майже дорівнює самій послідовності ($K(x) \approx |x|$).

Важливим теоретичним фундаментом є зв'язок між складністю та ймовірністю, що виражається через так звану універсальну ймовірність. Математично це співвідношення можна представити як:

$$K(x) = -\log_2 P(x) + O(1)$$

Це рівняння демонструє, що прості об'єкти (ті, що мають короткий код) мають значно вищу апріорну ймовірність появи, ніж складні. Цей результат став основою для принципу мінімальної довжини повідомлення (Minimum Message Length, MML) та принципу мінімальної довжини опису (Minimum Description Length, MDL) [2].

Принцип MDL стверджує: найкраща теоретична модель для заданого набору даних — це та модель, яка забезпечує найкоротший сумарний опис самих даних та параметрів самої моделі. У практичному

аналізі даних це дозволяє знаходити ідеальний баланс між точністю прогнозу та складністю математичного апарату.

Розглянемо основні сфери застосування цього підходу в сучасних ІТ:

1. *Машинне навчання*: Алгоритмічна складність допомагає боротися з проблемою «перенавчання» (overfitting). Якщо модель занадто складна, вона починає «заучувати» випадковий шум у навчальній вибірці. Використання критерію MDL змушує систему обирати більш прості, узагальнені гіпотези, які 2 краще працюють на нових даних.

2. *Стиснення даних*: Сучасні архіватори (LZMA, PPM тощо) фактично намагаються наблизитися до значення Колмогоровської складності. Коефіцієнт стиснення є прямим індикатором структурованості даних: якщо файл добре стискається, він містить багато закономірностей.

3. *Кібербезпека та виявлення аномалій*: Якщо певний потік мережевих даних раптово втрачає здатність до стиснення (його складність зростає до максимуму), це часто свідчить про наявність зашифрованого трафіку зловмисного ПЗ або проведення DDoS-атаки.

4. *Комп'ютерна лінгвістика*: Аналіз текстів через призму АТІ дозволяє оцінювати багатство мови, виявляти авторство або автоматично класифікувати тексти за стилем [1].

Незважаючи на елегантність, теорія Колмогорова стикається з бар'єром необчислюваності. Доведено, що не існує загального алгоритму, який для довільного рядка x видав би число $K(x)$. Це безпосередньо випливає з проблеми зупинки Тюрінга та теорем Геделя про неповноту.

Крім того, існує проблема залежності від мови опису. Хоча «Теорема інваріантності» стверджує, що для достатньо довгих рядків вибір мови програмування впливає лише на константну добавку, для малих обсягів даних цей вплив є критичним. На практиці вчені змушені замінювати істину складність $K(x)$ на «обчислювальну складність» або використовувати практичні алгоритми стиснення (наприклад, *Lempel – Ziv*), що є лише грубим наближенням ідеалу.

На нашу думку, основний розрив між теорією та практикою полягає в ігноруванні часового ресурсу. Колмогоровська складність не враховує, скільки часу знадобиться програмі для генерації об'єкта. У реальних

системах штучного інтелекту швидкість обробки часто важливіша за абсолютну мінімальність опису.

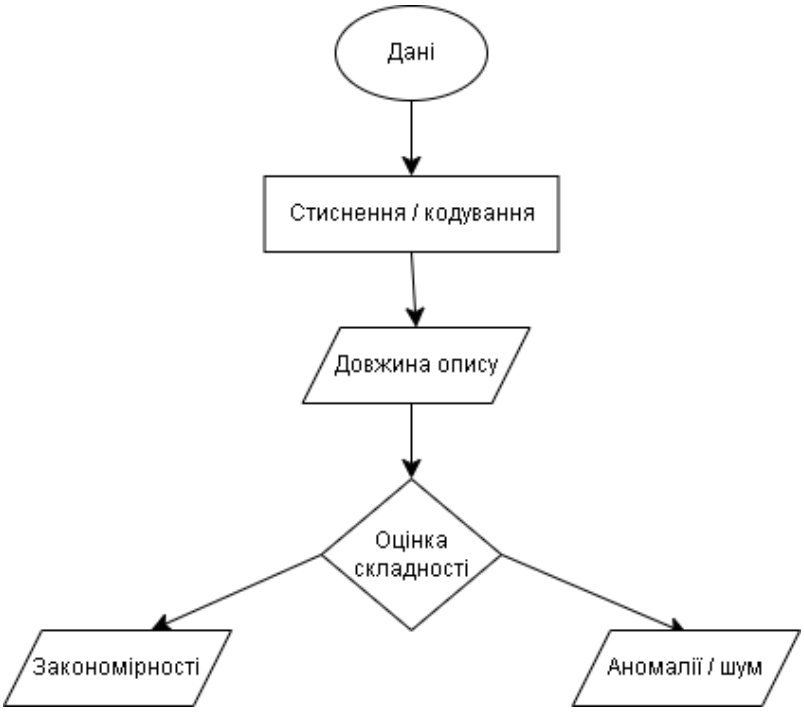

Рис. 1. Концептуальна схема зв'язку між даними та їх алгоритмічним описом

Висновок. Колмогоровська складність залишається «золотим стандартом» в оцінці інформаційної структури даних. Вона забезпечує фундамент для розуміння того, що таке випадковість і що таке закономірність. Хоча пряме обчислення K(x) є неможливим, похідні методи (MDL, стиснення) успішно працюють у системах розпізнавання образів та інтелектуального аналізу даних [3]. Подальші дослідження в цій галузі, ймовірно, будуть спрямовані на створення нових метрик, які б поєднували алгоритмічну стислість із 3 обчислювальною ефективністю, що стане ключем до створення більш досконалого штучного інтелекту.

Список використаних джерел

1. Льон О. В. Комп'ютерна лінгвістика: сучасне та майбутнє. Лінгвістичний портал MOVA.info. URL: http://www.mova.info/zbirnyk.pdf (дата звернення: 23.03.2026).

2. Учасники проектів Вікіпедіа. Мінімальна довжина повідомлення – Вікіпедія. Вікіпедія. URL:

https://uk.wikipedia.org/wiki/Мінімальна_довжина_повідомлення (дата звернення: 23.03.2026).

3. Grünwald P. D., Vitányi P. M. Algorithmic Information Theory. arXiv.org e-Print archive. URL: https://arxiv.org/pdf/0809.2754 (дата звернення: 23.03.2026).