№ 2(55) 2025

МЕТОДИ ТА ПРИЛАДИ КОНТРОЛЮ ЯКОСТІ

ISSN 1993-9981 print
ISSN 2415-3575 online

https://mpky.nung.edu.ua

Комп'ютерні технології та системи

Прийнято 02.10.2025. Прорецензовано 21.12.2025. Опубліковано 29.12.2025.

УДК 622.242.6, 681.5.015 DOI: 10.31471/1993-9981-2025-2(55)-178-189

ОЦІНКА ОСОБЛИВОСТЕЙ ЗАСТОСУВАННЯ КЛАСИФІКАЦІЙНО-ПОРІВНЯЛЬНОГО МЕТОДУ ОБРОБКИ КАРОТАЖНИХ ДАНИХ

Петришин Р. І.
Аспірант
Івано-Франківський національний технічний університет нафти і газу
76019, вул. Карпатська, 19, м. Івано-Франківськ, Україна
https://orcid.org/0009-0009-2770-6656
e-mail: romeopetryshyn@gmail.com

Мельник В. Д.*
Кандидат технічних наук, доцент
Івано-Франківський національний технічний університет нафти і газу
76019, вул. Карпатська, 19, м. Івано-Франківськ, Україна
https://orcid.org/0000-0001-7567-5625
e-mail: vitalii.melnyk@nung.edu.ua

Шекета В. І.
Доктор технічних наук, професор
Івано-Франківський національний технічний університет нафти і газу
76019, вул. Карпатська, 19, м. Івано-Франківськ, Україна
https://orcid.org/0000-0002-1318-4895
e-mail: vasyl.sheketa@nung.edu.ua

Халєєв Д. М.
Аспірант
Івано-Франківський національний технічний університет нафти і газу
76019, вул. Карпатська, 19, м. Івано-Франківськ, Україна
https://orcid.org/0009-0001-4548-231X
e-mail: dmytro.khalieiev-a174-23@nung.edu.ua

Запропоноване посилання: Петришин, Р. І., Пельгик, В. Д., Шекета, В. І., Халєєв, Д. М., Тріщ, В. В. & Богдан, О. Т. (2025). Оцінка особливостей застосування класифікаційно-порівняльного методу обробки каротажних даних. Методи та прилади контролю якості, 2(55), 178-189. doi: 10.31471/1993-9981-2025-2(55)-178-189

* Відповідальний автор

Тріщ В. В.
Аспірант
Івано-Франківський національний технічний університет нафти і газу
76019, вул. Карпатська, 19, м. Івано-Франківськ, Україна
https://orcid.org/0009-0000-8564-1752
e-mail: vlad.trishch@gmail.com

Богдан О. Т.
Аспірант
Івано-Франківський національний технічний університет нафти і газу
76019, вул. Карпатська, 19, м. Івано-Франківськ, Україна
https://orcid.org/0009-0000-9539-6359
e-mail: oleksiibohdan@gmail.com

Анотація. У нафтовидобувній галузі об'єктом дослідження є виявлення параметричної схожості та відмінностей між свердловинами з метою підвищення точності геологічного моделювання, оптимізації бурових робіт і прогнозування продуктивності пластів. У статті представлено модель для класифікації та порівняння каротажних даних свердловин на основі сучасних методів машинного та глибокого навчання. Наведено аналітичний огляд існуючих підходів до оцінювання кореляції між свердловинами, зокрема методів ручного аналізу, що базуються на експертних оцінках фахівців галузі, та використання простих коефіцієнтів подібності. Такі методи мають суттєві обмеження, пов'язані з високою суб'єктивністю, низькою відтворюваністю результатів і складністю масштабування на великі обсяги даних. Для подолання зазначених недоліків запропоновано підхід на основі глибокого навчання, який забезпечує автоматизацію процесів аналізу, підтримки та прийняття рішень у дослідженні свердловин. Розроблено модель, що використовує архітектуру рекурентної нейронної мережі (RNN), орієнтовану на обробку послідовних даних і виявлення довгострокових залежностей між ними, з використанням шарів нейронної мережі для оцінки подібності. Запропоновано візуалізацію на основі схем оцінки подібності для пари інтервалів та схему порівняльного навчання на основі триплетних втрат. Ефективність моделі оцінюється за допомогою метрик класифікації та кластеризації, що дозволяє кількісно визначити якість групування свердловин за параметрами подібності. Реалізований підхід є універсальним та надійним методом для впровадження процесу автоматизованого аналізу свердловин, спрямованого на зменшення залучення ручної праці фахівців і підвищення ефективності.

Ключові слова: аналіз свердловин, каротажні дані, глибоке навчання, нейронні мережі, моделювання, методологія, модель, агрегація даних, екстраполяція моделі, оцінка ефективності, автоматизація, прийняття рішень, кластеризація.

Вступ

У сучасному світі нафтогазова промисловість потребує постійного вдосконалення методів аналізу даних для підвищення ефективності видобутку та зменшення ризиків. Одним із ключових етапів є аналіз каротажних даних, що дозволяє отримати інформацію про геологічну будову родовища та властивості порід. Традиційні методи аналізу часто є трудомісткими та суб'єктивними, що призводить до зниження точності та ефективності прийняття рішень. Тому розробка автоматизованих методів аналізу каротажних даних є актуальною та важливою задачею.

У зв'язку з цим, застосування методів машинного (зокрема глибокого) навчання, стає все більш перспективним напрямком досліджень. Глибоке навчання дозволяє автоматизувати складні процеси аналізу даних, виявляти приховані закономірності та покращувати точність прогнозування. Застосування класифікаційно-порівняльних методів обробки каротажних даних на основі глибокого навчання дозволяє підвищити ефективність аналізу та прийняття рішень у нафтогазовій промисловості.

Мета роботи – розробка моделі класифікації та порівняння каротажних даних свердловин.

Аналіз сучасних закордонних і вітчизняних досліджень та публікацій

Виявлення подібності між свердловинами є необхідною частиною багатьох процедур. Це сприяє розумінню характеристик свердловини, плануванню стратегії

розробки родовища, виявленню аномалій під час процесу буріння, створенню синтетичних даних, наближених до реальних, і багатьом іншим задачам [1,2]. Сучасні методи оцінки кореляції між свердловинами в основному зосереджені на майже ручному аналізі характеристик експертами або оцінці деяких неприродних коефіцієнтів подібності, наприклад, простої евклідової відстані або більш складної ймовірності синхронізації [1]. Спільними для цих класичних підходів є відносно низька складність моделей і, відповідно, наявність обмежень для їх застосування. Вони вимагають складного вибору характеристик вхідних даних і гіперпараметрів моделі. Можливою альтернативою є рішення на основі даних. Такі моделі автоматизують прийняття рішень, зменшуючи кількість ручної праці, та менш схильні до помилок і суб'єктивності [2,3]. Однак, можна адаптувати потужні парадигми для навчання за подібністю [2,4]. Глибоке навчання і нейронні мережі (НМ) стали поширеним інструментом для розв'язання різних проблем. Це пов'язано зі здатністю нейронних мереж витягувати значущі представлення на основі даних і обчислювати подібності об'єктів на основі цих представлень.

Загальною проблемою для НМ є обмежені позначені дані. Проте, сучасні досягнення неконтрольованих і самоконтрольованих парадигм дозволяють дослідникам подолати цю проблему [5]. Ці методи автоматично виділяють ознаки або уявлення і, таким чином, працюють без конкретних суб'єктивних знань експертів.

Проаналізуємо наявні розробки в навчанні подібності, пов'язані з нафтогазовою промисловістю. Зосереджуємося на основних трьох напрямках досліджень, пов'язаних з простими підходами на основі правил, класичними підходами машинного навчання та підходами глибокого навчання. Підхід, заснований на правилах, пропонує визначити логічні правила для кореляції між свердловинами та визначення зон в інтервалах свердловин на основі попередніх знань експертів. Розглядають вирівнювання за допомогою динамічного викривлення часу на основі типових геометричних відстаней [6]. Після вирівнювання кривих експерт вирішує, чи вони схожі. Розроблено також метод оцінки подібності на основі статистичних методів [1]. Незважаючи на те, що згадані підходи досить прості для реалізації та легкі для інтерпретації, вони, в основному, зосереджені на розгляді однієї конкретної особливості добре реєстрованих даних та втрачають цінну інформацію. Крім того, більшість із них вимагає залучення експерта.

В останні роки підходи класичного машинного навчання і глибокого навчання досягли більше успіхів у розробці геологічних моделей. Використовують машину опорних векторів (SVM) для отримання плавного представлення даних, а потім застосовують загальні відстані кластеризації, як відстані Жаккара, для обчислення подібності між мітками свердловин [7]. Найбільш схожі свердловини використовуються для прогнозування каротажу. Хоча метод показує прийнятні результати, підтверджені оцінкою експертів, така модель відстані ігнорує природу даних, покладаючись на ретельно відібрані характеристики. Таким чином, метод не може ідентифікувати представлення свердловини та має обмежену застосовність, адаптовану до розглянутих діапазонів вибраних характеристик. Дослідження [2] запроваджує навчання подібності на основі даних вимірювання під час буріння (MWD) для виявлення аварій під час буріння. Автори демонструють застосування керованого навчання подібності з функціями, створеними експертами: вони агрегують статистику для часових інтервалів і вивчають класифікатор, який передбачає, чи відповідають інтервали різним або схожим аварійним випадкам. Очевидним обмеженням є ручне отримання мічених наборів даних для навчання такої системи, оскільки потрібні сотні мічених аварій.

Завдяки глибокому навчанню штучні нейронні мережі виконують кореляційну оцінку для свердловин. Ці моделі на основі даних використовують різні характеристики як вхідні дані: видобуток нафти, обводненість, тиск нагнітання, характеристики проникності або зображення, отримані з

цифрового каротажу. Нейронна мережа Convolution використовується для прогнозування регіональної проникності на основі набору геологічних особливостей [8].

Виклад основного матеріалу

Для прогнозування взаємозв'язку між видобувними свердловинами та оточуючими нагнітальними свердловинами на основі видобутку рідини та закачування води використовують мережу довготривалої короткочасної пам'яті. Після навчання мережі застосовують методи аналізу чутливості для отримання коефіцієнтів зв'язку між свердловинами.

Вся свердловина містить багато різних підшарів, тому, обчислюючи подібність між підінтервалами двох свердловин і потім об'єднуючи їх, можна дізнатися точніші моделі. Такий підхід є природним вибором, оскільки зазвичай свердловини мають різну довжину.

Розглядається набір даних, який складається з інтервалів свердловин; *i*-ий інтервал складається з відповідних складових: $X_i = \left(x_{1i}, ..., x_{l_i i}\right)$, $x_{ji} \in R^d$, де l_i є довжина *i*-го інтервалу, а d – це кількість ознак, зібраних на кожному кроці інтервалу.

Будуємо модель на основі даних, яка приймає як вхідні дані про два інтервали $\left(X_i, X_j\right)$ і виводить подібність між цими двома інтервалами $s_{ij} = s\left(X_i, X_j\right)$. Якщо інтервали близькі один до одного, модель подібності повинна виводити значення, близькі до 1, в іншому випадку оцінка схожості повинна бути близькою до *0*.

Розглядаються два підходи до позначення пар інтервалів для навчання класифікатора подібності:

Для *проблеми зв'язування* очікується виведення класифікатора 1, якщо два інтервали беруть з однієї свердловини, і 0, якщо із двох різних свердловин.

Для *проблеми закритого зв'язування* очікується виведення класифікатора 1, якщо два інтервали беруть з однієї свердловини і є близькі один до одного за глибиною, і 0, якщо вони не відповідають цим двом вимогам.

Двічі запускається рекурентна нейронна мережа (RNN), щоб обробити дані з двох інтервалів і отримати їх вкладення. Щоб оцінити подібність за парою вкладень, використовуються шари нейронної мережі. Під час навчання змінюються параметри RNN і частина оцінки подібності, щоб відповідати цільовим подібностям між парами інтервалів.

Розглядається трійка інтервальних об'єктів: якір, позитивний об'єкт, схожий на якір, і негативний об'єкт, несхожий на якір. Після вкладення триплетних інтервалів модель штрафується, якщо відстань між якорем та негативом менша за відстань між якорем та позитивом.

Через складність дані про нафту та газ вимагають ретельної попередньої обробки. Кроки в цьому контексті включають: 1) заповнення відсутніх значень; 2) роботу з конкретними особливостями каротажу; 3) вибір інтервалів зі свердловини, необхідних для навчання моделі. Потрібен додатковий крок для базових моделей, які потребують значущих ознак як вхідних даних: 4) генерація вектора ознак.

1) *Заповнення відсутніх значень.* Значна частка вимірювань у проаналізованих даних відсутня. Стратегія заповнення може дозволити використовувати більшість доступних даних. Якщо це можливо, заповнюємо кожне пропущене значення попереднім не пропущеним (пряме заповнення), інакше – найближчим доступним майбутнім значенням (зворотне заповнення). Оскільки використовується багато інтервалів для оцінки подібності між свердловинами, ці стратегії забезпечують достатнє покриття для роботи методів.

2) *Робота з особливостями.* Моделі машинного та глибокого навчання зазвичай вимагають обробки даних через специфіку технологічних умов у свердловинах під час каротажу. Попередня обробка включає такі кроки:

1. Виключення фізично неадекватних даних, тобто інтервалів, де питомий електричний опір менший або дорівнює 0.

2. Перетворення усіх даних питомого електричного опору в логарифмічний нормальний масштаб.

3. Нормалізація даних гамма каротажу в кожній свердловині та пласті шляхом віднімання середнього значення та розділення отриманих значень, щоб отримати одиничну дисперсію.

4. Нормалізація інших ознак шляхом віднімання середнього значення та ділення отриманих значень, щоб отримати одиничну дисперсію.

3) *Інтервали вибірки.* Вибираємо пари інтервалів фіксованої довжини 100 вимірювань зі свердловин для основних експериментів. Різниця в глибині між двома послідовними вимірюваннями становить 0,15 фута. Такі методи дозволяють створювати великі навчальні та перевірочні набори пар інтервалів, які можуть бути вирішальними для методів глибокого навчання. Для навчання формуємо стратифіковані вибірки, які мають майже однакову кількість пар подібних і різних інтервалів. Для інших частин експериментів, пов'язаних із побудовою представлення цілої свердловини, розділяємо свердловину на інтервали, що не перекриваються, однакової довжини l=100. У загальній класифікації стратегій вибірки можна назвати це стратифікованим варіантом пакетної вибірки для вивчення подібності.

4) *Генерація вектора ознак (якщо необхідно).* Оскільки розглядаються інтервальні рівні, то слід адаптувати дані для класичних методів машинного навчання. Типовим рішенням проблеми є агрегування ознак на інтервалах з використанням, наприклад, середнього та стандартного відхилень. Отже, кожен інтервал розміром (100, k) відповідає вектору ознак розміру 2k, де k – це кількість використаних значень. Перші k компоненти відповідають середнім значенням ознак, останні k компоненти відповідають стандартним відхиленням вздовж інтервалу.

Оскільки використовуються моделі глибокого навчання для обробки даних, інтервал трохи більшої або трохи меншої довжини також можна використовувати як вхідні дані для моделі подібності. Конкретна довжина 100 в цілому забезпечує прийнятну якість для побудованих моделей подібності, забезпечує швидку обробку та є достатньо великою, щоб фіксувати властивості на основі даних журналу. Для різних свердловин маємо різну довжину, тому вибірка дасть нам різну роздільну здатність для кожної свердловини, і відбирається однакова кількість інтервалів для кожної свердловини.

Для отримання подібності дотримуються загальної парадигми сучасного глибокого навчання на основі двофазної моделі з кодувальником на першому етапі та прийняттям рішень на другому етапі. Застосовується модель *кодера* $E_i = f\left(X_i\right)$, щоб отримати вкладення для кожного інтервалу. Потім порівнюється вбудовування за допомогою додаткової процедури $s_{ij} = g\left(E_i, E_j\right)$ з відсутністю або невеликою кількістю параметрів для оцінки під час навчання. Вихідні дані цієї частини повідомляють про подібність між інтервалами свердловин. На рисунку 1 представлена загальна схема того, як отримуються подібності між двома інтервалами, оціненими моделлю.

Існує два підходи до навчання таких моделей: один метод полягає в тому, щоб змусити модель вирішувати *проблему класифікації* (схожі об'єкти з пари чи ні); інша полягає в тому, щоб перемістити уявлення безпосередньо в латентному просторі.

Природною альтернативою для безпосередньої роботи з представленнями є використання *порівняльного навчання,* зображеного на рисунку 2. Розглянемо набір трійок. Кожен триплет складається з інтервалу прив'язки X_a, додатного інтервалу X_p і від'ємного інтервалу X_n. Їх вибирають так, щоб інтервал прив'язки та додатний інтервал були подібними за мірою подібності.

Формально триплетні втрати дорівнюють

$$\max\left(d\left(X_a, X_p\right) - d\left(X_a, X_n\right) - \alpha, 0\right), \quad (1)$$

де α – маржа: якщо $d\left(X_a, X_n\right)$ значно (більш ніж α) менше, ніж $d\left(X_a, X_p\right)$, то штрафується обчислена подібність і надалі модель повинна уникати таких ситуацій. Така втрата усереднюється для набору триплетів.

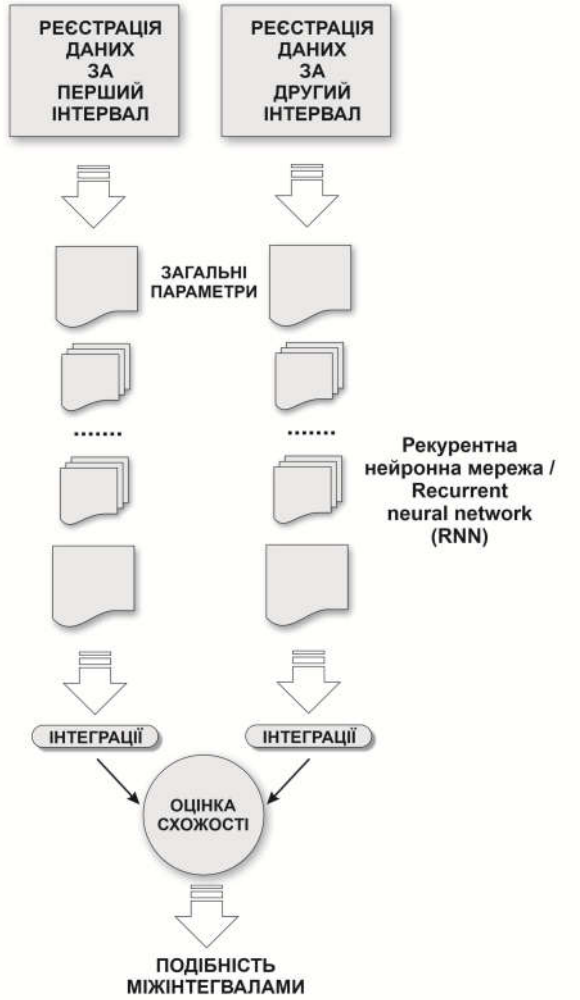

Рисунок 1 – Модель оцінки подібності для пари інтервалів

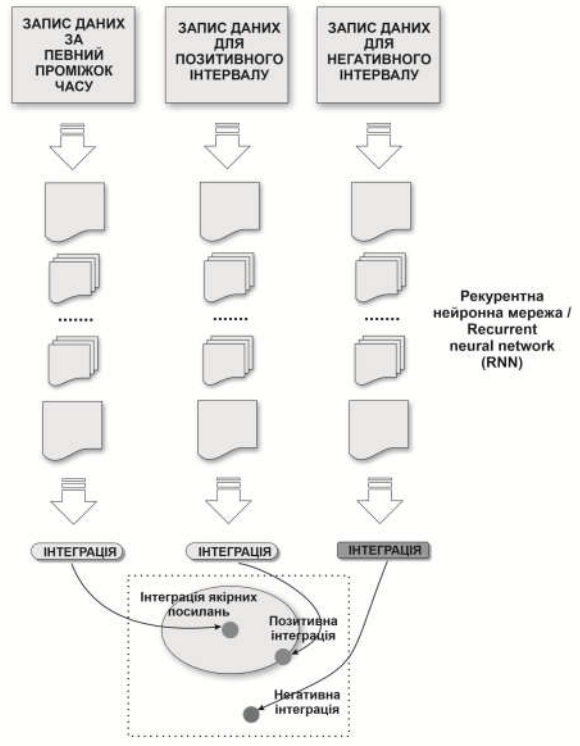

Рисунок 2 – Схема порівняльного навчання

На відміну від наведеного вище підходу на основі класифікації, *триплетна модель,* заснована на втраті триплетів, не вирішує безпосередньо проблеми зв'язування. Тому потрібен класифікатор над триплетною моделлю. Щоб обчислити подібності, оцінюємо евклідову відстань або косинусну відстань між отриманими вкладеннями з триплетної моделі, таким чином приймаючи аналогічну двофазну схему з кодувальником і прийняттям рішень на основі кодувальників.

Найбільш природним вибором для послідовних даних є архітектури моделей, засновані на рекурентних нейронних мережах, таких як рекурентна мережа з довгою короткочасною пам'яттю (LSTM) [9].

LSTM може фіксувати тимчасові залежності та обробляти довгострокові залежності, які з'являються в даних. Модель обробляє дані каротажу разом із її глибиною послідовно одне за одним спостереженням на певній глибині.

Проаналізуємо два варіанти оцінки схожості. Перший варіант полягає в тому, щоб обчислити евклідову (1) або косинусну (2) відстань, як у базовій лінії, але тепер між вбудовуваннями, щоб отримати подібність. Другий варіант полягає в тому, щоб навчити окрему повністю зв'язану (FC) нейронну мережу оцінювати подібність. Перевагою першого підходу є пряма оцінка відстані, тоді як у багатьох випадках додаткові повністю пов'язані проекційні шари дозволяють краще представити проміжні шари нейронної мережі.

Розглядаємо два підходи до навчання нейронних мереж: підхід на основі класифікації та підхід на основі порівняння. Основні стратегії агрегації подібностей інтервалів до подібності свердловин. Важливо перенести подібність від рівня інтервалу до рівня свердловини за допомогою належної стратегії агрегування, оскільки, зрештою, потрібні подібності між свердловинами. Щоб використовувати подібність на рівні інтервалу, реалізується наведена нижче процедура. Відокремлюємо для m-ої свердловини послідовні інтервали з даними X_j^m, $j = 1,...,k_m$. Те саме можна зробити

для n-ої свердловини. Для кожного такого інтервалу отримуємо вкладення $E_j^m = f_w\left(X_j^m\right) \in R^d$. У той же час для пари інтервалів $\left(X_j^m, X_i^m\right)$ модель подібності забезпечує оцінку подібності s_{ij}.

Застосовуємо макроагрегацію, якщо агрегуємо вкладення з інтервалу однієї свердловини, а потім отримуємо оцінку подібності для пари свердловин. Наприклад, використовуємо середнє значення для кожного компонента вектора вкладень:

$$E_j^m = \frac{1}{d}\sum_{j=1}^{k_m} E_{ij}^m. \qquad (2)$$

Потім оцінюємо подібність нашою моделлю $s_{mn} = g\left(E_m, E_n\right)$. Інший варіант – мікроагрегація, тобто спочатку обчислюємо подібності для пари інтервалів

$$s_{ij} = s\left(X_i^m, X_j^m\right), \qquad (3)$$

а потім агрегуємо до подібності свердловин

$$s_{mn} = \frac{1}{k_m k_n}\sum_{ij} s_{ij}. \qquad (4)$$

Щоб прискорити обчислення, оперуємо лише частиною можливих пар інтервалів, щоб отримати матрицю подібностей без використання надмірної суми. Агрегація через нейронну мережу. Досліджуємо модель, яка отримує дані каротажу для свердловин як вхідні дані (у вигляді послідовних інтервалів) і виводить вкладення всіх свердловин. Цей підхід є *наскрізним* навчанням. Частина, що відповідає вкладенню початкових даних реєстрації, є *модель кодера*, а частина, що відповідає агрегації вкладень інтервалів, є *модель агрегації*. Навчаємо агрегаційну модель для прогнозування маркування свердловин експертом. Запропонуємо етап наскрізної процедури навчання: здійснюється один крок оптимізації для моделі представлення інтервалів (будь-який із перерахованих вище); вибираються тренувальні свердловини з інтервалами відповідної довжини; введення інтервалів в отриману модель представлення, щоб отримати набір вкладень; введення послідовності вкладень в агрегаційну модель і навчання моделі.

Модель подібності повинна добре екстраполюватися, щоб бути ефективною. Розглядаємо дві властивості, пов'язані з екстраполяційною здатністю моделі: (1) застосовність у широкому діапазоні сценаріїв і (2) швидка адаптація до даних з нових областей за допомогою трансферного навчання або без нього.

Можливості адаптації моделі порівнюємо кількома поширеними способами:

1) Збереження моделі без змін і застосування її безпосередньо до нових даних, оскільки очікуємо, що нові дані будуть схожі на старі. Це найшвидший спосіб, який на практиці може погано працювати через відмінності в зборі та організації даних, а також через різні властивості різних нафтових родовищ.

2) Навчаємо нову модель з нуля, використовуючи лише нові дані. На відміну від попереднього, це найбільш небажаний підхід, який потребує багато часу та обчислювальних ресурсів. Крім того, якщо розмір нових даних обмежений, отримаємо модель із недостатньою продуктивністю.

3) Налаштовуємо стару модель, використовуючи нові дані, щоб зафіксувати особливості нових даних, що є компромісним рішенням.

Якщо незмінена або точно налаштована модель перевершує модель, навчену з нуля за допомогою нових даних, то отримуємо доказ того, що представлення та міри подібності є універсальними та можуть застосовуватися в різних сценаріях і для різних даних.

Робочий процес застосування моделі навчання подібності, що дозволяє вивчати моделі без міток, наданих експертами, складається з чотирьох кроків:

1) Попередня обробка: перетворює початкові необроблені дані та готує їх для використання як вхідних даних для кодувальника. Дуже важливо підтримувати подібний формат введення різноманітних даних, виключати нереалістичні вимірювання та продовжувати з відсутніми даними.

2) Вибірка: вибірка інтервалів дозволяє використовувати більше прикладів для навчання. Це також змушує модель вивчати шаблони меншого масштабу.

3) Обчислення подібності: отримуємо оцінки подібності між парами інтервалів за допомогою моделі на основі даних.

4) Сукупні оцінки подібності: повертаємося від рівня інтервалів до рівня свердловин шляхом агрегування оцінок, отриманих під час попереднього кроку обчислення.

Для проблеми класифікації модель на основі даних передбачає мітку класу для об'єкта. Модель намагається зробити якомога менше помилок, класифікуючи, наприклад, чи є пара інтервалів (об'єкт) подібною чи ні (дві можливі мітки класу). Зведені показники якості машинного навчання наведено в таблиці 1:

Таблиця 1 – Показники якості машинного навчання

Метрика	Задача	Діапазон
Точність	Класифікація	(0, 1)
ROC AUC	Класифікація	(0, 1)
PR AUC	Класифікація	(0, 1)
ARI	Кластеризація	(−1, 1)
AMI	Кластеризація	(−1, 1)
V-міра	Кластеризація	(0, 1)

Щоб оцінити якість моделей, придатних для проблем класифікації (наприклад, проблеми зв'язування), використовуються класичні показники машинного навчання: точність, площа під кривою робочих характеристик приймача (ROC AUC) і площа під кривою точності-відклику (PR AUC).

Розглядаємо набір даних $D=\left(X_i, y_i\right)$, $i=1,...,n$. У цьому позначенні $X_i \in R^{l\times 2d}$ є даними для пари інтервалів довжини l, що складається з *d*-значних мір на кожному кроці. Мітка y_i – це цільове значення. Для проблеми зв'язування y_i дорівнює 0 (різні інтервали) або 1 (подібні інтервали). Позначимо $\hat{y}_i = f\left(X_i\right)$ прогнозовану мітку моделі класифікатора $f\left(X\right)$.

Показники ROC AUC і PR AUC часто використовуються в машинному навчанні [2], оскільки краще підходять для загальної оцінки ефективності класифікатора.

Щоб визначити ROC AUC і PR AUC, спочатку вводимо більш базові показники, відображені в матриці збурення, що представляє ефективність моделі класифікації. Матриця збурення складається з чотирьох метрик: кількість об'єктів True Positive (TP), False Negative (FN), False Positive (FP), True Negative (TN):

$$TP = \frac{1}{N}\sum_{i=1}^{n}\left[y_i = 1\right]\left[\hat{y}_i = 1\right], \qquad (5)$$

$$FN = \frac{1}{N}\sum_{i=1}^{n}\left[y_i = 1\right]\left[\hat{y}_i = 0\right], \qquad (6)$$

$$FP = \frac{1}{N}\sum_{i=1}^{n}\left[y_i = 0\right]\left[\hat{y}_i = 1\right], \qquad (7)$$

$$TN = \frac{1}{N}\sum_{i=1}^{n}\left[y_i = 0\right]\left[\hat{y}_i = 0\right]. \qquad (8).$$

Ці показники відрізняють правильні мітки та помилки для першого та другого класу в проблемі класифікації.

Можна розглядати деякі пов'язані показники істинної позитивної частоти (TPR) або відкликання, помилкової позитивної частоти (FPR) і точності (PR):

$$TPR = \frac{TP}{TP + FN}, \qquad (9)$$

$$FPR = \frac{FP}{FP + TN}, \qquad (10)$$

$$PR = \frac{TP}{TP + FP}. \qquad (11)$$

Важливо зазначити, що, насправді, кожна модель класифікації передбачає не конкретну мітку, а ймовірності *p* приналежності до класу $p\left(X\right) = f_p\left(X\right) \in \left[0,1\right]$. Щоб отримати мітку, модель порівнює отримані ймовірності з деяким порогом p_0. Об'єкт належить до класу, якщо відповідна прогнозована ймовірність вище порогу $p > p_0$.

Порівнюючи отримані ймовірності з набором порогів від *0* до *1* і обчислюючи матриці збурення для кожного порогу,

отримуємо траєкторію в просторі (TPR, FPR) і подібну в просторі (TPR, PR) з кожною парою значень, що відповідають заданому порогу. Отримані криві називаються відповідно кривою ROC і кривою TPR-PR, а використовувані метрики – це площі під цими кривими.

Розглядаються три типи істинних цільових значень: вкладення інтервалів з однієї свердловини буде близьким один до одного. Для цього маркування не використовуються експертні оцінки; вкладення інтервалів з одного класу будуть близькі одне до одного, але вкладення можуть збігатися з різними свердловинами та лежати в різних шарах; вкладення інтервалів з одного класу та одного шару (позначення експерта) буде близьким одне до одного, але вбудовування можуть збігатися з різними свердловинами. Використовується скоригований індекс Ренда (ARI) як показник якості кластеризації, оскільки він забезпечує добре уявлення про якість кластеризації та має інтуїтивно зрозуміле пояснення. ARI вимірює подібність двох значень, ігноруючи перестановки порядку можливих міток. Формальне визначення ARI таке. Позначимо через n загальну кількість об'єктів з перевірочної вибірки, a – кількість пар в одному кластері з однаковою міткою, b – кількість пар, які знаходяться в різних кластерах і мають різні мітки. Отже, $a+b$ – кількість пар із подібним маркуванням за моделлю та відповідно до істинних міток. Тоді індекс Ренда (RI):

$$RI = \frac{a+b}{C_2^n}, \qquad (12)$$

де $C_2^n = \frac{n(n-1)}{2}$ – біноміальний коефіцієнт, який дорівнює загальній кількості різних пар для вибірки розміру n.

Скоригований індекс Ренда – це індекс Ренда, скоригований на ймовірність випадкового призначення міток:

$$ARI = \frac{RI - \mathrm{E}(RI)}{\max(RI) - \mathrm{E}(RI)}, \qquad (13)$$

де $\mathrm{E}(RI)$ є очікуванням індексу Ренда для випадкового вибору міток кластера.

Визначаємо дві альтернативні метрики: скориговану взаємну інформацію (AMI) і V-міру, які разом із ARI надають ширшу картину якості розглянутих моделей.

Визначимо дві мітки U і V та позначимо $P_i = \frac{|U_i|}{n}$, $P'_j = \frac{|V_j|}{n}$. Тоді ентропія U задається як:

$$H(U) = -\sum_{i=1}^{|U_i|} P_i \ln P_i. \qquad (14)$$

Подібним чином для V. Позначимо $P_{ij} = \frac{|U_i \cap V_j|}{n}$. Тоді взаємна інформація (MI) між U і V це:

$$MI(U,V) = P_{ij} \ln\left(\frac{P_{ij}}{P_i P'_j}\right). \qquad (15)$$

Скоригована взаємна інформація – це взаємна інформація, скоригована на можливість випадкового призначення міток:

$$AMI = \frac{MI - \mathrm{E}[MI]}{mean(H(U), H(V)) - \mathrm{E}[MI]}, \qquad (16)$$

де $\mathrm{E}[MI]$ – очікувана взаємна інформація для випадкового вибору.

Розглянемо V-міру, яка є гармонійним середнім однорідності h і повноти c. Однорідність – це властивість розподілу кластерів, що кожен кластер містить лише члени одного класу:

$$h = 1 - \frac{H(C|K)}{H(C)}. \qquad (17)$$

Повнота – це властивість розподілу кластера, згідно з якою всі члени даного класу призначаються одному кластеру:

$$h = 1 - \frac{H(K|C)}{H(K)}. \qquad (18)$$

$H(C|K)$ є умовною ентропією класів із заданими кластерними призначеннями і $H(C)$ є ентропією класів:

$$H(C|K) = -\sum_{c=1}^{|C|}\sum_{k=1}^{|K|}\frac{n_{ck}}{n}\ln\left(\frac{n_{ck}}{n}\right), \quad (19)$$

де n_{ck} є кількість об'єктів *c*-го кластера, які потрапляють в *k*-ий клас.

$$H(C) = -\sum_{c=1}^{|C|}\frac{n_c}{n}\ln\left(\frac{n_c}{n}\right). \quad (20)$$

Тоді *V*-міра:

$$v = \frac{2hc}{h+c}. \quad (21)$$

Висновки

Запропоновано комплексну оцінку ефективності моделі для класифікації та порівняння каротажних даних, що включає використання двох груп метрик. Використано метрики класифікації, такі як ROC AUC та PR AUC, для визначення належності інтервалів до однієї свердловини, що дає змогу оцінити здатність моделі правильно ідентифікувати інтервали, що відносяться до однієї свердловини. Для оцінки якості створених моделлю представлень застосовано метрики кластеризації Adjusted Rand Index (ARI), Adjusted Mutual Information (AMI) та V-міра, що дозволяє об'єктивно оцінити якість кластеризації результатів та порівняти їх з експертними оцінками. Особливу увагу приділено дослідженню властивостей запропонованої моделі до екстраполяції, що робить можливою роботу з новими базами даних без необхідності додаткового навчання та перенаалаштування моделі під нові умови. Реалізований підхід є універсальним та надійним методом для реалізації процесу автоматизованого аналізу свердловин в контексті зменшення залучення ручної праці фахівців.

Подяки

Відсутні.

Конфлікт інтересів

Відсутній.

Список використаних джерел / References

1. Verma A.K., Routray A., Mohanty W.K. Assessment of similarity between well logs using synchronization measures. *IEEE Geoscience and Remote Sensing Letters.* 2014. Vol.11. No.12. pp. 2032-2036. doi: 10.1109/LGRS.2014.2317498

2. Akkurt R., Conroy T.T., Psaila D., Paxton A., Low J., Spaans P. Accelerating and enhancing petrophysical analysis with machine learning: a case study of an automated system for well log outlier detection and reconstruction. SPWLA 59th Annual Logging Symposium, Society of Petrophysicists and Well-Log Analysts. 2018.

3. Koroteev D., Tekic Z. Artificial intelligence in oil and gas upstream: Trends, challenges, and scenarios for the future. *Energy AI.* 2021. Vol. 3. Article 100041.doi: 10.1016/j.egyai.2020.100041

4. Brazell S., Bayeh A., Ashby M., Burton D. A machine-learning-based approach to assistive well-log correlation. *Petrophysics*-the SPWLA J. Form. Eval. Reserv. Descr. 2019. Vol. 60(04). pp. 469-479. doi:10.30632/PJV60N4-2019A1

5. Jing L., Tian Y. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell*. 2020. Vol. 43. No 11. pp. 4037-4058. doi: 10.1109/TPAMI.2020.2992393

6. Zoraster S., Paruchuri R., Darby S. Curve alignment for well-to-well log correlation. SPE Annual Technical Conference and Exhibition, OnePetro. 2004. doi: 10.2118/90471-MS

7. Ali M., Jiang R., Ma H., Pan H., Abbas K., Ashraf U., Ullah J. Machine learning-a novel approach of well logs similarity based on synchronization measures to predict shear sonic logs *J. Pet. Sci. Eng*. 2021. Vol. 203.Article 108602.doi: 10.1016/j.petrol.2021.108602

8. Du S., Wang R., Wei C., Wang Y., Zhou Y., Wang J., Song H. The connectivity evaluation among wells in reservoir utilizing machine learning methods. *IEEE Access*.2020. Vol. 8. pp. 47209-47219.doi: 10.1109/ACCESS.2020.2976910

9. Greff K., Srivastava R.K., Koutník J., Steunebrink B.R., Schmidhuber J. LSTM: A search space odyssey. *IEEE Trans. Neural Netw. Learn. Syst.*. 2016. Vol. 28. No 10. pp. 2222-2232. doi: 10.48550/arXiv.1503.04069

ASSESSMENT OF PECULIARITIES OF APPLYING THE CLASSIFICATION-COMPARATIVE METHOD OF PROCESSING WELL LOGGING DATA

Petryshyn R. I.
Postgraduate Student
Ivano-Frankivsk National Technical University of Oil and Gas
76019, Karpatska Street, 19, Ivano-Frankivsk, Ukraine
https://orcid.org/0009-0009-2770-6656
e-mail: romeopetryshyn@gmail.com

Melnyk V. D.
Candidate of Technical Sciences, Associate Professor
Ivano-Frankivsk National Technical University of Oil and Gas
76019, Karpatska Street, 19, Ivano-Frankivsk, Ukraine
https://orcid.org/0000-0001-7567-5625
e-mail: vitalii.melnyk@nung.edu.ua

Sheketa V. I.
Doctor of Technical Sciences, Professor
Ivano-Frankivsk National Technical University of Oil and Gas
76019, Karpatska Street, 19, Ivano-Frankivsk, Ukraine
https://orcid.org/0000-0002-1318-4895
e-mail: vasyl.sheketa@nung.edu.ua

Khaleiev D. M.
Postgraduate Student
Ivano-Frankivsk National Technical University of Oil and Gas
76019, Karpatska Street, 19, Ivano-Frankivsk, Ukraine
https://orcid.org/0009-0001-4548-231X
e-mail: dmytro.khalieiev-a174-23@nung.edu.ua

Trishch V. V.
Postgraduate Student
Ivano-Frankivsk National Technical University of Oil and Gas
76019, Karpatska Street, 19, Ivano-Frankivsk, Ukraine
https://orcid.org/0009-0000-8564-1752
e-mail: vlad.trishch@gmail.com

Bohdan O. T.
Postgraduate Student
Ivano-Frankivsk National Technical University of Oil and Gas
76019, Karpatska Street, 19, Ivano-Frankivsk, Ukraine
https://orcid.org/0009-0000-9539-6359
e-mail: oleksiibohdan@gmail.com

Abstract. In the oil production industry, the research objective is to identify parametric similarities and differences between wells to improve the accuracy of geological modeling, optimize drilling operations, and predict reservoir productivity. The article presents a model for the classification and comparison of well logging data based on modern machine learning and deep learning methods. An analytical review of existing approaches for evaluating correlation between wells is provided, including manual analysis based on expert assessments and the use of simple similarity coefficients. These methods have signi-?icant limitations associated with high subjectivity, low reproducibility of results, and difficulties in scaling to large datasets. To overcome these drawbacks, a deep learning–based approach is proposed that enables automation of data analysis, decision support, and pattern recognition in well studies. The developed model employs a recurrent neural network (RNN) architecture designed to process sequential data and capture long-term dependencies, utilizing neural network layers to assess similarity between well intervals. A visualization method based on similarity evaluation schemes for interval pairs and a comparative learning approach using a triplet loss function are proposed. The model's performance is evaluated using classification and clustering metrics, which allow for quantitative assessment of well grouping quality according to similarity parameters. The implemented approach represents a universal and reliable method for automating well data analysis, aimed at reducing manual expert involvement and improving overall operational efficiency in geological and production processes.

Keywords: well analysis, well logging data, deep learning, neural networks, modeling, methodology, model, data aggregation, model extrapolation, effectiveness evaluation, automation, decision-making, clustering.