УДК 004.89 **doi: 10.26906/SUNZ.2024.3.158**

Е. Г. Фастовський[1], А. І. Роговий[1], О. Б. Ахієзер[1], А. В. Фролов[2], Р. В. Артюх[2, 3]

[1] Національний технічний університет "Харківський політехнічний інститут", Харків, Україна
[2] Харківський національний університет радіоелектроніки, Харків, Україна
[3] ДП "Південний державний ПК і НД інститут авіаційної промисловості", Харків, Україна

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ АНАЛІЗУ ТА СИНТЕЗУ ПОЯСНЕННИХ МОДЕЛЕЙ ШТУЧНОГО ІНТЕЛЕКТУ НА ОСНОВІ ВЕРБАЛЬНИХ МЕТОДІВ

Анотація. Предметом дослідження є аналіз та синтез поясненних моделей штучного інтелекту. **Мета** роботи - розробка інформаційної технології аналізу та синтезу поясненних моделей штучного інтелекту на основі вербальних методів. У статті вирішуються такі **завдання**: аналіз математичних формул та методів, що використовуються для пояснення рішень, що приймають моделі штучного інтелекту, аналіз методів, класів, фреймворків та функцій програмних бібліотек, а також їх використання для пояснення рішень, що приймають моделі штучного інтелекту, синтез поясненних вербальних моделей штучного інтелекту, розробка проєкту системи синтезу поясненних вербальних моделей штучного інтелекту. Використовуються такі **методи**: системний аналіз, вербальні методи прийняття рішень (формування системи понять визначеної предметної галузі, формування порядкової класифікації станів об'єкта/процесу, впорядкування станів об'єкта/процесу з певного класу, визначення найкращого стану об'єкта/процесу), методи моделювання та проєктування інформаційних систем (діаграми варіантів використання, діаграми діяльності). Здобуто такі **результати**: проведено аналіз математичних формул та методів, що використовуються для пояснення рішень, що приймають моделі штучного інтелекту. Запропоновано підхід до синтезу поясненних вербальних моделей штучного інтелекту. Проведено аналіз методів, класів, фреймворків та функцій програмних бібліотек, а також їх використання для пояснення рішень, що приймають моделі штучного інтелекту. Розроблено проєкт системи синтезу поясненних вербальних моделей штучного інтелекту. **Висновки**: методи вербального аналізу виявляються ефективними для синтезу поясненних моделей штучного інтелекту, що включає кілька етапів: визначення системи понять, створення критеріальних описів станів, їх класифікація, впорядкування та обрання найкращого стану. Вони підкреслюють важливість використання лінгвістичної інформації разом з числовими даними для комплексного аналізу складних проблем. Інтегруючи елементи вербального аналізу в пояснені моделі штучного інтелекту, можна покращити взаємодію з користувачем, його розуміння і сприйняття систем штучного інтелекту.

Ключові слова: поясненні моделі штучного інтелекту; вербальні методи прийняття рішень.

Вступ

Розробленим моделям штучного інтелекту (ШІ) сьогодні бракує можливості пояснення у багатьох важливих галузях їх застосування. Наприклад, у банківській сфері, коли модель ШІ відхиляє заявку на отримання кредиту, важливо пояснити заявнику причини відхилення та які коригувальні дії він може вжити, щоб отримати кредит. Або у медичній діагностиці, коли модель ШІ прогнозує розвиток діабету у людини, важливо пояснити чому та які фактори сприятимуть розвитку захворювання в майбутньому. Нарешті, при експлуатації автономного транспортного засобу, коли модель ШІ ідентифікує об'єкти на дорозі, важливо пояснити чому було прийнято певне рішення щодо руху цього транспорту в певній ситуації. Моделі ШІ на основі машинного або глибокого навчання розв'язують задачу пошуку функції, що задовольняє як навчальним, так і тестовим даним. Але функціональний зв'язок зазвичай настільки складний, що його не можна пояснити кінцевому користувачу моделі ШІ. Тому доречно застосувати для пояснення вербальні методи, що враховують можливості та обмеження людини в процесі обробки інформації при вирішенні складних слабо структурованих проблем, що описуються якісними факторами.

Аналіз проблеми й наявних методів

Коли рішення повністю приймаються машинами, а люди завжди на стороні одержувача, виникає гостра необхідність зрозуміти, як машини дійшли цих рішень. Моделі, на яких засновані системи ШІ, часто називають моделями "чорної скриньки". Отже, існує необхідність у зрозумілості та інтерпретованості моделей для того, щоб пояснити зроблені ними прогнози [1].

Лінійні моделі. Лінійні моделі, такі як лінійна регресія для прогнозування реальної величини виходу або модель логістичної регресії для передбачення класу та відповідних ймовірностей, є алгоритмами контрольованого навчання. Ці лінійні моделі для контрольованого машинного навчання є дуже простими для інтерпретації. Їх також легко пояснити зацікавленим особам [2–5].

Лінійна регресія використовується для прогнозування кількісного результату цільової змінної з огляду на набір предикторов. Формула моделювання зазвичай виглядає так:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon. \qquad (1)$$

Бета-коефіцієнти відомі як параметри, а епсилон-член відомий як похибка. Похибку можна розглядати як консолідований показник, який відображає нездатність моделі прогнозувати. Нічого не можна передбачити зі 100% точністю в реальному світі, оскільки варіації даних є реальністю. Дані постійно змінюються. Метою розробки моделі є прогнозування з якомога можливою точністю та стабільністю. Цільова змінна приймає значення члена перехоплення, коли незалежні змінні приймають нульове значення.

Модель лінійної регресії застосовується, коли цільова ознака є безперервною. Коли ж цільова ознака є

двійковою, як-от 0 або 1, істина чи хибність, прийняти чи відхилити, то модель лінійної регресії не застосовується. Це пояснюється тим, що прогнозоване значення для цільової функції може перевищувати діапазон 0 і 1, але очікується обмеження вихідних даних двома класами, оскільки потрібно передбачити два класи окремо. Ось чому потрібна модель логістичної регресії, яка використовує двійкові значення для обчислення логарифмічних шансів, відомих як відношення шансів. Співвідношення шансів лінійно пов'язане з характеристиками. Саме тому модель логістичної регресії відома як лінійна модель [6–9].

Модель логістичної регресії можна пояснити за допомогою наступного рівняння:

$$Sigmoid(t) = \frac{1}{1+e^{-t}}. \quad (2)$$

Наведена вище функція також відома як сигмоїдна функція.

$$ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n. \quad (3)$$

$ln\left(\frac{P}{1-P}\right)$ – логарифмічна ймовірність результату. Бета-коефіцієнти, зазначені у наведеному вище рівнянні, пояснюють, як ймовірність змінної результату збільшується або зменшується на кожну одиницю збільшення або зменшення пояснюючої змінної. Форма сигмоїдної функції виглядає як s-подібна крива.

Інтерпретація моделі логістичної регресії істотно відрізняється від моделі лінійної регресії. Зважена сума з правої частини рівняння перетворюється на значення ймовірності. Ліва частина рівняння називається логарифмічною ймовірністю, оскільки це логарифм відношення ймовірності того, що подія відбудеться, до ймовірності того, що не відбудеться.

Нелінійні моделі. Дерево рішень – це нелінійна модель, яка пов'язує незалежну змінну із залежною. На локальному рівні це можна розглядати як кусково-лінійну регресію, але на глобальному рівні – це нелінійна модель, оскільки немає однозначного зв'язку між залежною та незалежними змінними. На відміну від лінійної регресійної моделі немає математичного рівняння, що показує взаємозв'язок між вхідними та вихідними змінними. Якщо зберігати параметр максимальної глибини дерева на нескінченному рівні, то дерево рішень може ідеально підходити до даних, що є класичним сценарієм надлишкового припасування моделі. Незалежно від того, чи можна навчальний набір даних лінійно поділити чи ні, дерева рішень схильні до надмірного припасування. Із цим необхідно боротися. Зазвичай вдаються до обрізки дерев, щоб отримати найкращу модель. Можна розглядати дерево рішень як послідовність умовних тверджень, результатом яких є значення або клас у вихідному стовпчику. Наприклад, якщо людині 45 років, вона працює у приватному секторі і має ступінь доктора філософії, то вона безперечно заробляє понад 50 тис. дол. на рік. Дерева рішень – це алгоритм контрольованого навчання, який застосовується, коли існує безліч можливих комбінацій характеристик, здатних вплинути на цільовий стовпець. У дереві рішень генеральна сукупність розбивається на дві чи більше підгруп з урахуванням найбільш значущого роздільника чи диференціатора у вхідних змінних [10-13].

Ансамблеві моделі. Ансамблеві моделі – це найскладніший набір моделей, які потребують детального пояснення, оскільки результат є об’єднаним результатом кількох передбачень. Ансамбль має на увазі просто групування. Що важливо у випадку ансамблевих моделей, так це те, як пояснити прогнози, який варіант моделі фактично створив прогноз, і як прочитати граничний внесок функцій у остаточний процес прогнозування [14].

Перевага моделі дерева рішень полягає в тому, що вона враховує потенційну нелінійність, яка існує в наборі даних. Змінні взаємодії вступають у гру під час створення прогнозів моделі. Однак обмеження моделі дерева рішень полягає в тому, що вона схильна до упередженості, оскільки потужні або сильніші функції беруть участь у процесі побудови дерева, а слабкі функції не можуть увійти в процес розгалуження дерева, оскільки їм бракує передбачуваної сили. Таким чином, модель стає упередженою до кількох вибіркових і сильніших характеристик із набору даних. Це також іноді призводить до переобладнання моделі. Щоб збалансувати вплив сильних функцій, важливо регулювати введення функцій у моделі на основі дерева. Якщо залишити сильні функції поза етапом створення моделі та включити лише слабкі функції до створення дерева, все одно можна генерувати прогнози, але це знову буде упереджена модель. Таким чином, лікування зміщення моделі та водночас контроль переобладнання можна зробити лише тоді, коли використовують комбінацію сильних і слабких функцій у процесі побудови дерева. Комбінацію можна виконати виключно на основі методу завантаження, а кількість дерев можна збільшити достатньо, щоб усереднити прогнози [15].

Ансамблеві моделі бувають трьох типів: моделі упаковки (bagging, також відомі як моделі агрегування початкового завантаження), моделі підвищення (boosting) та моделі укладання (stacking). Моделі укладання можуть бути двох видів: групове укладання (same group stacking) та компонування різних варіантів (different variant model stacking). Групове укладання включає однорідні типи моделей, такі як тільки включення деревоподібних моделей і зіставлення кожного результату моделі з іншими моделями. Гетерогенна модель укладання означає накладення деревоподібних та недеревоподібних моделей одна на одну та поєднання їх прогнозів. Зрозумілість моделей укладання зовсім не складна, оскільки можна ідентифікувати конкретну модель і пояснити прогнози та параметри. Більш складними є моделі упаковки та підвищення. Модель випадкового лісу є прикладом моделі упаковки, де вирощується багато дерев, та їх прогнози поєднуються для досягнення кінцевого результату [16]. Моделі підвищення беруть базову модель, навчаються на її результатах та намагаються покращити модель ітеративним способом [17].

Моделі часових рядів. Основною метою моделі часових рядів є оцінка значення цільової змінної з використанням часу як незалежної змінної. Цільовою змінною може бути вартість акцій, кількість одиниць

продукту, сума доходу, який надходитиме на рахунок компанії або кількість унікальних відвідувачів певного веб-сайту. Прогнозовані значення є багатокроковими, оскільки, використовуючи модель часових рядів, зазвичай, прогнозують кілька часових кроків. Модель часових рядів генерує прогнозні значення. Вони мають певні рівні довіри. Чим вищий рівень достовірності, тим краща модель. На нижчих рівнях довіри моделі бракує стабільності у створенні прогнозних значень. Довірчий інтервал можна розрахувати як прогнозоване значення очікуваного значення плюс-мінус 1,96 (стандартизоване значення зі статистичної таблиці, що відповідає 95% довіри), помножене на стандартну помилку залишкового терміну, розрахованого на основі моделі. Це засновано на умові нормального розподілу члена помилки.

Модель часових рядів вимагає, щоб дані записувалися через часті проміжки часу, без будь-яких перерв у часовому кроці. Дані часових рядів є впорядкованими за своєю природою, оскільки порядок визначає неявну послідовність часу. Інженеру з машинного навчання важливо створювати корисні функції з даних, щоб робити правильні прогнози. У моделі часових рядів час є незалежною змінною. В однофакторній моделі часових рядів є лише одна змінна. У моделі причинного прогнозування використовують модель, подібну до регресійної моделі. В однофакторній моделі часових рядів ознаками є авторегресійні терміни, такі як терміни з відставанням, терміни ковзного середнього (наприклад, триперіодне або п’ятиперіодне ковзне середнє) і різницеві терміни. Найпопулярніші моделі прогнозування часових рядів спираються виключно на історичні значення цільової змінної [18-22].

Моделі обробки природної мови. Сьогодні весь світ пов'язаний між собою Всесвітнім павутинням (World Wide Web). Неструктуровані текстові дані трапляються всюди. Вони знаходяться в соціальних мережах, у розмовах електронною поштою, у чатах через різні програми, на HTML-сторінках, у текстових документах, у службах підтримки клієнтів, відповідях на онлайн-опитування та багато іншого. Нижче наведено деякі варіанти використання неструктурованих даних:

– класифікація документів, коли на вході знаходиться текстовий документ, а виходом може бути бінарний клас або багатокласовий ярлик;

– іноді якщо настрої позначені, вони також відповідають сценарію класифікації документів. Інакше класифікація настроїв буде класифікацією з урахуванням лексикону;

– моделі розпізнавання іменованих сутностей (named entity recognition – NER), де на вході перебуває текстовий документ, але в виході – клас іменованих сутностей;

– оцінка якості текстового опису задач спринту, ідентифікації та класифікація їх на «зрозумілі» та «не зрозумілі»;

– резюме тексту, коли великий текст стиснутий і представлений в компактній формі.

У класифікації текстів звичайними є моделі NER і передбачення наступного слова, де вхід – це пропозиція чи вектори слів, а вихід – мітка, яку потрібно класифікувати чи передбачити. До настання ери машинного навчання завдання класифікації тексту вирішувалася вручну, коли група анотаторів читала, розуміла зміст тексту, що є у документі, і відносила його до певного класу. З масштабним зростанням обсягу неструктурованих даних ручний спосіб класифікації став дуже складним. Тепер деякі анотовані дані можуть бути введені в комп'ютер і застосований алгоритм навчання, щоб навчену модель можна було в майбутньому використовувати для прогнозування.

Неструктуровані документи чи вхідні текстові вектори дуже багатовимірні. Прогностичні моделі, що використовуються для класифікації документів, повинні бути пояснені, тому що причини, що лежать в основі прогнозу, або особливості, що лежать в основі передбаченого класу, мають бути показані кінцевому користувачеві. У випадку класифікації тексту модель передбачає клас 1 проти класу 2, тому важливо знати ключові слова, які є позитивними та негативними для класу 1. У багатокласовій класифікації це стає більш складним, оскільки необхідно пояснити всі ключові слова, які призводять до передбачення певного класу [23-29].

Існують різні інструменти та механізми для створення зрозумілості моделей. Бібліотеки Python з відкритим вихідним кодом мають деякі переваги та недоліки [30].

SHAP. Бібліотека SHAP (SHapley Additive exPlanations) – це уніфікований підхід на базі Python для пояснення результатів будь-якої моделі машинного навчання. Бібліотека SHAP Python заснована на теорії ігор із локальними поясненнями. Підхід теорії ігор – це спосіб отримати прогнози за наявності одного чинника проти його відсутністю. Якщо відбувається значна зміна в очікуваному результаті, то фактор є дуже важливим для цільової змінної. Цей метод поєднує кілька попередніх методів для пояснення результатів, що генеруються моделями машинного навчання. Фреймворк SHAP може бути використаний для різних типів моделей, крім моделей на основі часових рядів. Бібліотека SHAP може бути використана для осмислення моделей [31].

LIME. LIME розшифровується як локальні інтерпретовані пояснення, що не залежать від моделі (Local Interpretable Model-Agnostic Explanations). Локальне відноситься до пояснення локальності класу, що був передбачений моделлю. Поведінка класифікатора при локальності дає розуміння прогнозу. Інтерпретованість означає, що коли передбачення не може бути інтерпретовано людиною, то в ньому немає сенсу. Отже, передбачення класів мають бути інтерпретованими. Незалежність від моделі передбачає, що замість розуміння конкретного типу моделі система та метод мають бути здатними генерувати інтерпретації.

Проблема класифікації тексту (наприклад, аналіз настрою) – коли на вхід подаються тексти документів, а на виході отримується клас. Коли модель передбачає позитивний настрій для тексту, необхідно знати, які слова змусили модель передбачити клас як

позитивний. Ці вектори слів іноді дуже прості, наприклад, окремі слова. Іноді вони складні (наприклад, вкраплення слів), і в цьому випадку потрібно знати, як модель інтерпретувала вкраплення слів і як це впливає на класифікацію. У цих сценаріях бібліотека LIME є надзвичайно корисною для розуміння сенсу моделей машинного навчання та глибокого навчання [32-35].

ELI5. ELI5 – це бібліотека на базі Python, призначена для створення зрозумілого конвеєра ШІ, який дозволяє візуалізувати та налагоджувати різні моделі машинного навчання за допомогою уніфікованого API. Вона має вбудовану підтримку кількох фреймворків та надає можливість пояснювати "чорні ящики" моделей. Мета бібліотеки – зробити пояснення простими для всіх видів моделей "чорного ящику" [36-42].

Мета статті - проаналізувати математичні формули та методи, що використовуються для пояснення рішень, яks приймають моделі штучного інтелекту; провести аналіз методів, класів, фреймворків та функцій програмних бібліотек, а також їх використання для пояснення рішень, що приймають моделі ШІ; запропонувати підхід до синтезу поясненних вербальних моделей штучного інтелекту; розробити проєкт системи синтезу поясненних вербальних моделей штучного інтелекту.

Матеріали та методи дослідження

Інформаційна технологія аналізу та синтезу пояснювальних моделей штучного інтелекту на основі вербальних методів передбачає використання обробки природної мови й машинного навчання для створення розмовних систем ШІ, які можуть розуміти людську мову та реагувати на неї. Ці системи можна використовувати в різних додатках, таких як чатботи, віртуальні асистенти та голосові помічники, для покращення розуміння рекомендації, що надаються користувачеві [1, 43, 44].

Методи вербального аналізу використовуються для розв'язування складних неструктурованих проблем. Дослідження в цій галузі зосереджені на розробці методів підтримки прийняття рішень, які включають як числові, так і вербальні аспекти. Вони звертають увагу на важливість включення вербальних елементів у процеси прийняття рішень, підкреслюючи цінність лінгвістичної інформації поряд з числовими даними. Такий підхід дозволяє проводити більш комплексний аналіз складних проблем, які не можуть бути повністю охоплені лише кількісними методами [45]. Ці розробки мають значення для навчання та аналізу комунікативних патернів за допомогою технологій ШІ [46].

Синтез пояснювальних вербальних моделей ШІ починається з формування системи понять певної предметної галузі. Виявляються критерії, якими користуються фахівці при аналізі об'єктів або процесів в певній предметній галузі. Для кожного критерію формується шкала впорядкованих значень: від кращих до гірших або від більш характерних для об'єкта/процесу в певній ситуації до менш характерних. Це дозволяє формувати критеріальні описи всіх гіпотетично можливих станів об'єкта/процесу, порівнювати їх через використання відповідних графів домінування станів, що будуються на основі шкал критеріїв, та визначати й пояснювати чому певний стан об'єкта/процесу кращий/гірший, ніж інший [47].

На другому етапі відбувається формування порядкової класифікації станів об'єкта/процесу. Визначається назва та лінійний порядок класів. Класифікація відбувається за принципами "кращій стан не може потрапити у гірший клас" та "гірший стан не може потрапити у кращій клас". Це дозволяє сформувати повну класифікацію через аналіз обмеженої кількості формально визначених найінформативніших станів, їх пряму та непряму класифікацію кращих/гірших станів, яка обмежує для них відповідні класи. Через це обмеження можна контролювати класифікацію на наявність протиріч та пояснювати, чому певний стан об'єкта/процесу належить до визначеного класу. Особливе значення мають граничні стани кожного класу, для яких зміна значення за певним критерієм призводить до переходу об'єкта/процесу в кращій/гірший клас. Це дозволяє визначати та пояснювати ефективність певної траєкторії покращення стану об'єкта/процесу через зміни значень за певними критеріями.

На третьому етапі обирається певний клас станів об'єкта/процесу та формується їх впорядкування. Для цього будуються єдині порядкові шкали для всіх можливих пар критеріїв, які потім об'єднуються в єдину порядкову шкалу всіх критеріїв, що використовується для порівняння станів об'єкта/процесу з визначеного класу, які неможливо було порівняти через відповідні графи домінування станів. Можливість побудови єдиної порядкової шкали всіх критеріїв без циклів є ознакою відсутності протиріч впорядкування станів, а її використання пояснює чому певний стан об'єкта/процесу кращий/гірший, ніж інший.

На четвертому етапі відбувається вибір найкращого стану серед тих, що не вдалося впорядкувати на попередньому етапі. Для цього обираються кращі стани з кожної впорядкованої гілки. Випадковим чином обирається пара станів для порівняння та визначаються їх відносні недоліки за кожним критерієм. Ці недоліки впорядковуються за важливістю для користувача та відбувається процес їх взаємної компенсації. За результатами порівняння визначається кращий стан з цієї пари, який порівнюється з іншим випадковим станом за такою ж процедурою. Наприкінці залишається тільки один найкращій стан об'єкта/процесу.

Отже, методи вербального аналізу виявляються ефективними для синтезу поясненних моделей ШІ, що включає кілька етапів: визначення системи понять; створення критеріальних описів станів; їх класифікація, впорядкування та обрання найкращого стану. Вони підкреслюють важливість використання лінгвістичної інформації разом з числовими даними для комплексного аналізу складних проблем.

Для розробки системи синтезу пояснених моделей штучного інтелекту (ШІ) необхідно врахувати кілька ключових аспектів [48–50]:

– зіставлення вербальної поведінки людини з поведінкою системи ШІ, враховуючи особистісні

критерії, щоб генерувати відповідні вербальні та невербальні комбіновані моделі поведінки;

– покращення пояснення систем ШІ в користувацьких інтерфейсах, зосередившись на зрозумілості, персоналізації та візуалізації пояснень, визнанні компромісів, усуненні потенційних помилкових уявлень і прив'язці пояснень до ментальних моделей користувачів задля прозорості та сприйняття користувачами рекомендацій систем ШІ;

– перевірка результатів дослідження в реальних умовах через вивчення персоналізованих пояснень, візуалізацію та забезпечення відображення пояснень у користувацькому інтерфейсі.

Вербальний аналіз рішень [45] – це ефективний інструмент, який може вдосконалити системи пояснюваного ШІ, надаючи словесні пояснення моделей машинного навчання та міркувань за допомогою слів, тексту або природної мови. Наприклад, цей підхід може допомогти людям краще зрозуміти процес прийняття рішень системами ШІ і потенційно зменшити кількість пацієнтів, яким необхідно пройти тривалий і багатоетапний процес діагностики [51].

Синтез пояснювальних вербальних моделей ШІ передбачає чотири основні етапи [52] (рис. 1):

– формування системи понять визначеної предметної галузі;

– формування порядкової класифікації станів об’єкта/процесу;

– впорядкування станів об’єкта/процесу з певного класу;

– визначення найкращого стану об’єкта/процесу.

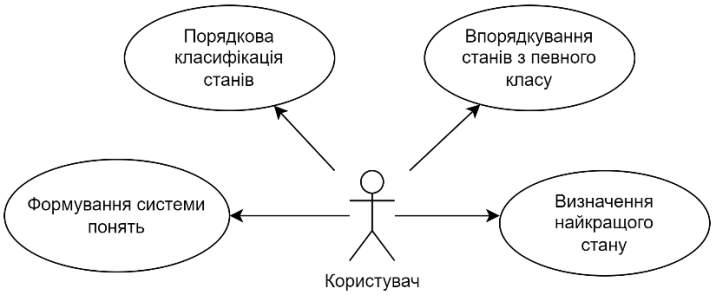

Рис. 1. Діаграма варіантів використання системи

Формування системи понять визначеної предметної галузі складається з таких діяльностей (рис. 2):

– виявлення критеріїв порівняння об’єктів/процесів певної предметної галузі;

– формування порядкових шкал для критеріїв;

– побудова графа домінування станів об’єктів/процесів;

– порівняння визначених станів та пояснення типу відношення між ними (краще, гірше, непорівнянні).

Порядкова класифікація гіпотетично можливих станів складається з таких діяльностей (рис. 3):

– визначення назв та лінійного порядку класів;

– визначення найінформативнішого стану;

– пряма класифікація найінформативнішого стану;

– перевірка несуперечливості прямої класифікації;

– непряма класифікація станів;

– класифікація реальних станів;

– визначення та пояснення ефективності певної траєкторії покращення стану.

Впорядкування станів об’єкта/процесу з певного класу складається з таких діяльностей (рис. 4):

– побудова єдиної порядкової шкали (ЄПШ) для всіх пар критеріїв;

– побудова єдиної порядкової шкали для всіх критеріїв;

– перевірка несуперечливості ЄПШ;

– порівняння визначених станів та пояснення типу відношення між ними (краще, гірше, непорівнянні).

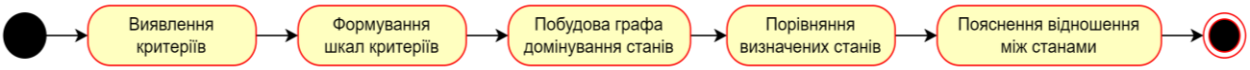

Рис. 2. Діаграма діяльностей формування системи понять

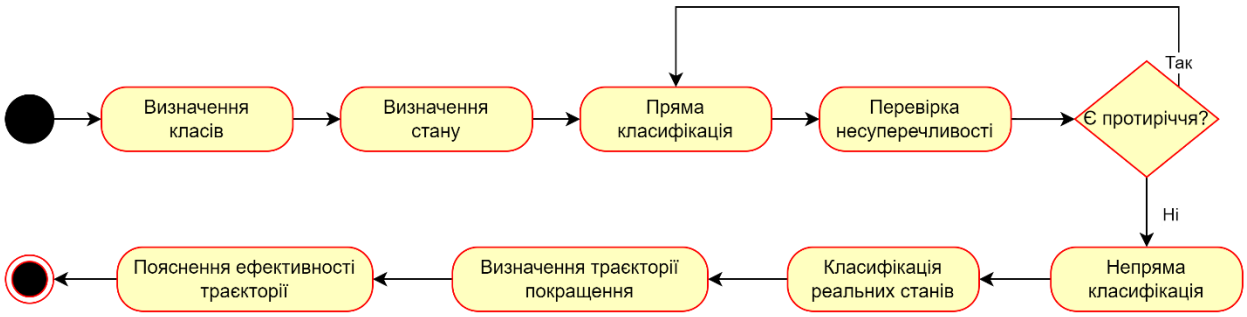

Рис. 3. Діаграма діяльностей порядкової класифікації станів

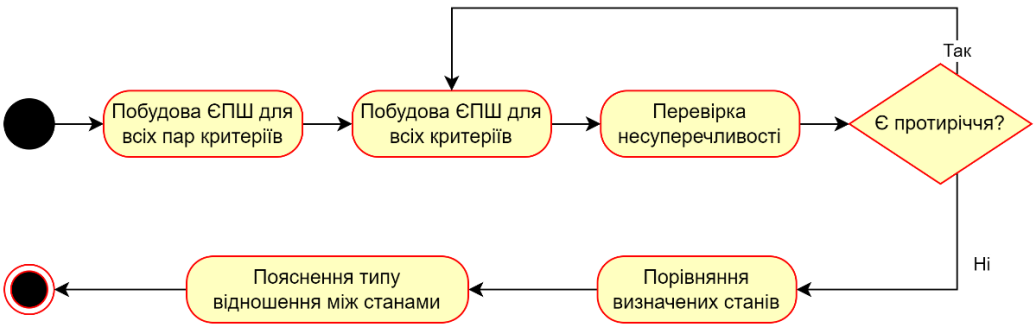

Рис. 4. Діаграма діяльностей впорядкування станів

Визначення найкращого стану об'єкта/процесу складається з таких діяльностей:

– вибір пари станів для порівняння;

– визначення відносних недоліків станів за кожним критерієм;

– впорядкування недоліків за важливістю для користувача;

– взаємна компенсація недоліків;

– визначення кращого стану.

Результати

Проведено аналіз математичних формул та методів, що використовуються для пояснення рішень, що приймають моделі штучного інтелекту:

– лінійні моделі;

– нелінійні моделі;

– ансамблеві моделі;

– моделі часових рядів;

– моделі обробки природної мови.

Проведено аналіз методів, класів, фреймворків та функцій програмних бібліотек, а також їх використання для пояснення рішень, що приймають моделі штучного інтелекту:

– SHAP;

– LIME;

– ELI5.

Запропоновано підхід до синтезу поясненних вербальних моделей штучного інтелекту.

Розроблено проєкт системи синтезу поясненних вербальних моделей штучного інтелекту.

Висновки

Отже, методи вербального аналізу виявляються ефективними для синтезу поясненних моделей штучного інтелекту, що включає кілька етапів: визначення системи понять, створення критеріальних описів станів, їх класифікація, впорядкування та обрання найкращого стану.

Вони підкреслюють важливість використання лінгвістичної інформації разом з числовими даними для комплексного аналізу складних проблем.

Інтегруючи елементи вербального аналізу в пояснені моделі штучного інтелекту, можна покращити взаємодію з користувачем, його розуміння і сприйняття систем штучного інтелекту.

Список літератури

1. Mishra, P. Model Explainability and Interpretability. *Practical Explainable AI Using Python*. Apress, Berkeley, CA. 2022. P. 1–22. DOI: https://doi.org/10.1007/978-1-4842-7158-2_1
2. Montesinos López, O. A., Montesinos López, A., Crossa, J. Linear Mixed Models. *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. Springer, Cham. 2022. P. 141–170. DOI: https://doi.org/10.1007/978-3-030-89010-0_5
3. Starbuck, C. Linear Regression. *The Fundamentals of People Analytics.* Springer, Cham. 2023. P. 181–206. DOI: https://doi.org/10.1007/978-3-031-28674-2_10
4. Elements of Generalized Linear Mixed Models / Salinas Ruíz, J. et al. *Generalized Linear Mixed Models with Applications in Agriculture and Biology.* Springer, Cham. 2023. P. 1–42. DOI: https://doi.org/10.1007/978-3-031-32800-8_1
5. Linear Regression / James, G. et al. *An Introduction to Statistical Learning. Springer Texts in Statistics.* Springer, Cham. 2023. P. 69–134. DOI: https://doi.org/10.1007/978-3-031-38747-0_3
6. Wüthrich, M. V., Merz, M. Generalized Linear Models. *Statistical Foundations of Actuarial Learning and its Applications. Springer Actuarial*. Springer, Cham. 2023. P.111–205. DOI: https://doi.org/10.1007/978-3-031-12409-9_5
7. Kleinbaum, D. G., Klein, M. Important Special Cases of the Logistic Model. *Logistic Regression. Statistics for Biology and Health*. Springer, New York, NY. 2010. P. 41–71. DOI: https://doi.org/10.1007/978-1-4419-1742-3_2
8. Logistic Regression / Weisburd, D. et al. *Advanced Statistics in Criminology and Criminal Justice*. Springer, Cham. 2022. P. 127–185. DOI: https://doi.org/10.1007/978-3-030-67738-1_4
9. Logistic Regression / Backhaus, K. et al. *Multivariate Analysis. Springer Gabler*, Wiesbaden. 2021. P. 267–354. DOI: https://doi.org/10.1007/978-3-658-32589-3_5
10. Grąbczewski, K. Techniques of Decision Tree Induction. *Meta-Learning in Decision Tree Induction. Studies in Computational Intelligence.* 2014. Vol. 498. P. 11–117. DOI: https://doi.org/10.1007/978-3-319-00960-5_2
11. Barros, R.C., de Carvalho, A.C.P.L.F., Freitas, A.A. Decision-Tree Induction. *Automatic Design of Decision-Tree Induction Algorithms. SpringerBriefs in Computer Science*. Springer, Cham. 2015. P. 7–45. DOI: https://doi.org/10.1007/978-3-319-14231-9_2
12. Kozak, J. Ant Colony Decision Tree Approach. *Decision Tree and Ensemble Learning Based on Ant Colony Optimization. Studies in Computational Intelligence.* 2019. Vol. 781. P. 45–80. DOI: https://doi.org/10.1007/978-3-319-93752-6_3

13. Construction of Optimal Decision Trees and Deriving Decision Rules from Them / Azad, M. et al. *Decision Trees with Hypotheses. Synthesis Lectures on Intelligent Technologies*. Springer, Cham. 2022. P. 41–53. DOI: https://doi.org/10.1007/978-3-031-08585-7_4
14. Seni, G., Elder, J. F. Importance Sampling and the Classic Ensemble Methods. *Ensemble Methods in Data Mining. Synthesis Lectures on Data Mining and Knowledge Discovery*. Springer, Cham. 2010. P. 53–87. DOI: https://doi.org/10.1007/978-3-031-01899-2_4
15. Seni, G., Elder, J. F. Predictive Learning and Decision Trees. *Ensemble Methods in Data Mining. Synthesis Lectures on Data Mining and Knowledge Discovery*. Springer, Cham. 2010. P. 14–27. DOI: https://doi.org/10.1007/978-3-031-01899-2_2
16. Cutler, A., Cutler, D. R., Stevens, J. R. Random Forests. *Zhang, C., Ma, Y. (eds) Ensemble Machine Learning*. Springer, New York, NY. 2012. P. 157–175. DOI: https://doi.org/10.1007/978-1-4419-9326-7_5
17. Ferreira, A. J., Figueiredo, M. A. T. Boosting Algorithms: A Review of Methods, Theory, and Applications. *Zhang, C., Ma, Y. (eds) Ensemble Machine Learning. Springer*, New York, NY. 2012. P. 35–85. DOI: https://doi.org/10.1007/978-1-4419-9326-7_2
18. Štulajter, F. Random Processes and Time Series. *Predictions in Time Series Using Regression Models*. Springer, New York, NY. 2002. P. 51–71. DOI: https://doi.org/10.1007/978-1-4757-3629-8_2
19. Das, M., Ghosh, S. K. Standard Bayesian Network Models for Spatial Time Series Prediction. *Enhanced Bayesian Network Models for Spatial Time Series Prediction. Studies in Computational Intelligence*. 2020. Vol. 858. P. 11–22. DOI: https://doi.org/10.1007/978-3-030-27749-9_2
20. Estimation Under Normal Mixture Models for Financial Time Series Data /Sun, L H. Et al. *Copula-Based Markov Models for Time Series. SpringerBriefs in Statistics()*. Springer, Singapore. 2020. P. 55–72. DOI: https://doi.org/10.1007/978-981-15-4998-4_4
21. Brockwell, P. J., Davis, R. A. ARMA Models. *Introduction to Time Series and Forecasting. Springer Texts in Statistics*. Springer, Cham. 2016. P. 73–96. DOI: https://doi.org/10.1007/978-3-319-29854-2_3
22. Interval Evaluation of Stationary State Probabilities for Markov Set-Chain Models / O. Akhiiezer et al. *10th International Conference on Advanced Computer Information Technologies (ACIT)*. 2020. P. 82–85. DOI: https://doi.org/10.1109/ACIT49673.2020.9208932
23. Petrov, S. Latent Variable Grammars for Natural Language Parsing. *Coarse-to-Fine Natural Language Processing. Theory and Applications of Natural Language Processing*. Springer, Berlin, Heidelberg. 2011. P. 7–46. DOI: https://doi.org/10.1007/978-3-642-22743-1_2
24. Goyal, P., Pandey, S., Jain, K. Introduction to Natural Language Processing and Deep Learning. *Deep Learning for Natural Language Processing*. Apress, Berkeley, CA. 2018. P. 1–74. DOI: https://doi.org/10.1007/978-1-4842-3685-7_1
25. Kulkarni, A., Shivananda, A. Advanced Natural Language Processing. *Natural Language Processing Recipes*. Apress, Berkeley, CA. 2019. P. 97–128. DOI: https://doi.org/10.1007/978-1-4842-4267-4_4
26. Harris, I. G., Harris, C. B. Generation of Verification Artifacts from Natural Language Descriptions. *Soeken, M., Drechsler, R. (eds) Natural Language Processing for Electronic Design Automation*. Springer, Cham. 2020. P. 37–70. DOI: https://doi.org/10.1007/978-3-030-52273-5_3
27. Catta, D., Moot, R., Retoré, C. Dialogical Argumentation and Textual Entailment. *Loukanova, R. (eds) Natural Language Processing in Artificial Intelligence—NLPinAI 2020. Studies in Computational Intelligence*. 2021. Vol. 939. P. 191–226. DOI: https://doi.org/10.1007/978-3-030-63787-3_7
28. Rodrigues, M., Teixeira, A. Extracting Relevant Information Using a Given Semantic. *Advanced Applications of Natural Language Processing for Performing Information Extraction. SpringerBriefs in Electrical and Computer Engineering()*. Springer, Cham. 2015. P. 37–50. DOI: https://doi.org/10.1007/978-3-319-15563-0_4
29. Гринченко, М., Роговий, М. Модель ідентифікації задач спринту проєкту на основі їх опису. Сучасний стан наукових досліджень та технологій в промисловості, 2023. 4(26), P. 33–44. https://doi.org/10.30837/ITSSI.2023.26.033
30. Gianfagna, L., Di Cecco, A. Explainable AI: Needs, Opportunities, and Challenges. *Explainable AI with Python*. Springer, Cham. 2021. P. 27–46. DOI: https://doi.org/10.1007/978-3-030-68640-6_2
31. Mishra, P. Explainability for Ensemble Supervised Models. *Explainable AI Recipes*. Apress, Berkeley, CA. 2023. P. 119–206. DOI: https://doi.org/10.1007/978-1-4842-9029-3_4
32. Handling Missing Values in Local Post-hoc Explainability / Cinquini, M. et al. *Longo, L. (eds) Explainable Artificial Intelligence. xAI 2023. Communications in Computer and Information Science.* 2023. Vol. 1902. P. 256–278. DOI: https://doi.org/10.1007/978-3-031-44067-0_14
33. Explainable AI Methods - A Brief Overview / Holzinger, A. et al. *Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, KR., Samek, W. (eds) xxAI - Beyond Explainable AI. xxAI 2020. Lecture Notes in Computer Science()*. 2022. Vol. 13200. P. 13–38. DOI: https://doi.org/10.1007/978-3-031-04083-2_2
34. Kamath, U., Liu, J. Post-Hoc Interpretability and Explanations. *Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning*. Springer, Cham. 2021. P. 167–216. DOI: https://doi.org/10.1007/978-3-030-83356-5_5
35. Kumar, D., Mehta, M. A. An Overview of Explainable AI Methods, Forms and Frameworks. *Mehta, M., Palade , V., Chatterjee, I. (eds) Explainable AI: Foundations, Methodologies and Applications. Intelligent Systems Reference Library.* 2023. Vol. 232. 2023. P. 43–59. DOI: https://doi.org/10.1007/978-3-031-12807-3_3
36. Vuppalapati, C. ML Models: Food Security and Climate Change. *Artificial Intelligence and Heuristics for Enhanced Food Security. International Series in Operations Research & Management Science*. 2022. Vol. 331. P. 395–518. DOI: https://doi.org/10.1007/978-3-031-08743-1_6
37. Sonar Signal Prediction Using Explainable AI for IoT Environment / Yadav, T. et al. *Bansal, H.O., Ajmera, P.K., Joshi, S., Bansal, R.C., Shekhar, C. (eds) Next Generation Systems and Networks. BITS-EEE-CON 2022. Lecture Notes in Networks and Systems.* 2023. Vol. 641. P. 209–222. DOI: https://doi.org/10.1007/978-981-99-0483-9_19
38. Brad, S., Ștetco, E. An Interactive Artificial Intelligence System for Inventive Problem-Solving. *Nowak, R., Chrząszcz, J., Brad, S. (eds) Systematic Innovation Partnerships with Artificial Intelligence and Information Technology. TFC 2022. IFIP Advances in Information and Communication Technology*. 2022. Vol. 655. P. 165–177. DOI: https://doi.org/10.1007/978-3-031-17288-5_15

39. Yan, W. Q. Convolutional Neural Networks and Recurrent Neural Networks. *Computational Methods for Deep Learning. Texts in Computer Science*. Springer, Singapore. 2023. P. 69–124. DOI: https://doi.org/10.1007/978-981-99-4823-9_3
40. Paaß, G., Giesselbach, S. Foundation Models for Text Generation. *Foundation Models for Natural Language Processing. Artificial Intelligence: Foundations, Theory, and Algorithms*. Springer, Cham. 2023. P. 227–311. DOI: https://doi.org/10.1007/978-3-031-23190-2_6
41. Vuppalapati, C. The Role of the Government and the AI Readiness. *Machine Learning and Artificial Intelligence for Agricultural Economics. International Series in Operations Research & Management Science*. 2021. Vol. 314. P. 479–548. DOI: https://doi.org/10.1007/978-3-030-77485-1_7
42. Ye, A., Wang, Z. Data Preparation and Engineering. *Modern Deep Learning for Tabular Data*. Apress, Berkeley, CA. 2023. P. 95–179. DOI: https://doi.org/10.1007/978-1-4842-8692-0_2
43. Фастовський Е. Г, Єльчанінов Д. Б. Інформаційна технологія аналізу та синтезу поясненних моделей штучного інтелекту. *Теоретичні та практичні дослідження молодих вчених [Електронний ресурс] : зб. тез доп. 17-ї Міжнар. наук.-практ. конф. магістрантів та аспірантів, 28-30 листопада 2023 р.* С. 85. URL: https://repository.kpi.kharkov.ua/handle/KhPI-Press/71485
44. Conversational AI: What Is It, How Does It Work, and Why Does It Matter?. URL: https://www.247.ai/insights/conversational-ai-what-it-and-how-does-it-work (дата звернення: 29.02.2024)
45. Moshkovich H, Mechitov A, Olson D. Verbal Decision Analysis. *Multiple Criteria Decision Analysis: State of the Art Surveys. International Series in Operations Research & Management Science*. 2005. Vol. 78. P. 609–633. DOI: https://doi.org/10.1007/0-387-23081-5_15
46. Butow P, Hoque E. Using artificial intelligence to analyse and teach communication in healthcare. *The Breast*. 2020. Vol. 50; P. 49–55. DOI: https://doi.org/10.1016/j.breast.2020.01.008
47. Фастовський Е. Г. Синтез поясненних вербальних моделей штучного інтелекту. *Інформаційні технології: теорія і практика. I (VII) міжнародна науково-практична конференція здобувачів вищої освіти і молодих учених "Інформаційні технології: теорія і практика". Тези доповідей (Дніпро 20 – 22 березня 2024)*. 2024. С. 225–227, URL: https://ir.nmu.org.ua/handle/123456789/166564
48. AKM Bahalul Haque, A. K. M. Najmul Islam, Patrick Mikalef. Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. *Technological Forecasting & Social Change*. 2023. Vol. 186, Part A. Article 122120. DOI: https://doi.org/10.1016/j.techfore.2022.122120
49. How to explain AI systems to end users: a systematic literature review and research agenda / Laato S. et al. *Internet Research*. 2022. Vol. 32, No 7. P. 1–31.
50. Aly A., Tapus A. A model for synthesizing a combined verbal and nonverbal behavior based on personality traits in human-robot interaction. *8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 2013. P. 325–332. DOI: https://doi.org/10.1109/HRI.2013.6483606
51. A Protocol for the Diagnosis of Autism Spectrum Disorder Structured in Machine Learning and Verbal Decision Analysis / Evandro Andrade et al. *Computational and Mathematical Methods in Medicine*. 2021. Vol. 2021. Article ID 1628959. DOI: https://doi.org/10.1155/2021/1628959
52. Фастовський Е. Г. Проєкт системи синтезу поясненних вербальних моделей штучного інтелекту. *28-й Міжнародний молодіжний форум "Радіоелектроніка та молодь у XXI столітті". Зб. матеріалів форуму (Харків 16 – 18 квітня 2024)*. 2024. Т. 6. С. 74–76. DOI: https://doi.org/10.30837/IYF.IIS.2024.074

Received (Надійшла) 08.04.2024
Accepted for publication (Прийнята до друку) 19.06.2024

Information technology of analysis and synthesis of explained models of artificial intelligence based on verbal methods

Eduard Fastovskyi, Anton Rogovyi, Olena Akhiiezer, Andrii Frolov, Roman Artiukh

Abstract. The **subject matters** of the article is the analysis and synthesis of explained models of artificial intelligence. The **goal** of the work is to develop an information technology for analyzing and synthesizing explained models of artificial intelligence based on verbal methods. The following **tasks** were solved in the article: analysis of mathematical formulas and methods used to explain decisions made by artificial intelligence models, analysis of methods, classes, frameworks, and functions of software libraries, as well as their use to explain decisions made by artificial intelligence models, synthesis of explained verbal models of artificial intelligence, development of a system for synthesizing explained verbal models of artificial intelligence. The following **methods** are used: system analysis, verbal decision-making methods (formation of a system of concepts in a particular subject area, formation of an ordinal classification of object/process states, ordering object/process states from a particular class, determining the best state of an object/process), methods of modeling and designing information systems (use case diagrams, activity diagrams). The following **results** were obtained: the mathematical formulas and methods used to explain decisions made by artificial intelligence models are analyzed. An approach to the synthesis of explained verbal models of artificial intelligence is proposed. The methods, classes, frameworks, and functions of software libraries, as well as their use to explain decisions made by artificial intelligence models, are analyzed. A project of a system for synthesizing explained verbal models of artificial intelligence was developed. **Conclusions**: verbal analysis methods prove to be effective for synthesizing explained artificial intelligence models, which includes several stages: defining a system of concepts, creating criterion descriptions of states, classifying them, organizing them, and selecting the best state. They emphasize the importance of using linguistic information together with numerical data for a comprehensive analysis of complex problems. By integrating elements of verbal analysis into explained artificial intelligence models, it is possible to improve user interaction, understanding, and perception of artificial intelligence systems.

Keywords: explained models of artificial intelligence; verbal decision-making methods.