DOI: 10.20998/2079-0023.2024.01.16
УДК 004.8:004.9

С. Ф.ЧАЛИЙ, доктор технічних наук, професор, Харківський національний університет радіоелектроніки, професор кафедри інформаційних управляючих систем, м. Харків, Україна; e mail: serhii.chalyi@nure.ua, ORCID: https://orcid.org/0000-0002-9982-9091

В. О. ЛЕЩИНСЬКИЙ, кандидат технічних наук (PhD), доцент, Харківський національний університет радіоелектроніки, доцент кафедри програмної інженерії; м. Харків, Україна, ; volodymyr.leshchynskyi@nure.ua, ORCID: https://orcid.org/0000-0002-8690-5702

ПОБУДОВА МОЖЛИВІСНИХ ПРИЧИННО-НАСЛІДКОВИХ ЗАЛЕЖНОСТЕЙ МІЖ КЛАСАМИ ЕКВІВАЛЕНТНОСТІ ДАНИХ В ІНТЕЛЕКТУАЛЬНІЙ ІНФОРМАЦІЙНІЙ СИСТЕМІ

Предметом дослідження є процеси формування пояснень щодо прийняття рішень в системі штучного інтелекту. Пояснення в таких системах дають можливість зробити прозорим та зрозумілим процес формування рішень для користувача і, як наслідок, підвищити довіру користувача до отриманих результатів. Мета роботи полягає у розробці підходу до побудови ймовірнісної каузальної моделі пояснення з урахуванням класів еквівалентності вхідних, проміжних і результуючих даних. Вирішення цієї задачі створює умови для побудови пояснень у формі причинно-наслідкових залежностей на основі доступної інформації про властивості вхідних даних, а також про властивості отриманих у системі штучного інтелекту результатів. Для досягнення мети вирішуються такі задачі: розробка моделі каузальної залежності між класами еквівалентності вхідних та вихідних даних; розробка методів побудови класів еквівалентності даних процесу прийняття рішень та методу побудови причинно-наслідкового представлення пояснення. Запропоновано ймовірнісну модель каузальної залежності, що містить причинно-наслідковий зв'язок між класами еквівалентності вхідних або проміжних та результуючих даних, отриманих у процесі прийняття рішень в системі штучного інтелекту. Цей зв'язок враховує оцінки можливості і необхідності такої залежності. Модель створює умови для пояснення можливих причин отриманого рішення. Запропоновано комплекс методів побудови класів еквівалентності даних процесу прийняття рішень та побудови причинно-наслідкового представлення пояснення, що встановлює каузальний зв'язок між класами еквівалентності. При побудові класів еквівалентності встановлюються відношення обов'язкового і необов'язкового уточнення даних, вимоги або виключення даних, а також кон'юнкції даних. При побудові причинно-наслідкового представлення пояснення розраховується можливість та обмеження необхідності такої залежності, що дає можливість побудувати пояснення на основі доступної інформації про отримані рішення та вхідні і проміжні дані, які були використані для формування цих рішень.

Ключові слова: каузальна залежність, причинно-наслідкова залежність, темпоральна залежність, можливість, необхідність, пояснення, система штучного інтелекту, інтелектуальна система, зрозумілий штучний інтелект, інформаційна система.

Вступ. Ефективне використання сучасних інтелектуальних інформаційних систем базується на застосуванні методів машинного навчання. Останні імплементують алгоритми прийняття рішень, що використовують виявлені в наборах даних залежності [1]. Однак складні навчені моделі, як правило, є непрозорими і, як наслідок, незрозумілими для користувачів систем штучного інтелекту, що приводить до зниження довіри до отриманих рішень. Така невідповідність може привести до відмови користувачів від використання рішень, наданих системою штучного інтелекту, або ж до менш ефективного застосування цих рішень [2, 3]. Для вирішення даної проблеми користувачам надаються пояснення, що представляють причини прийнятих рішень, а також окремих дій, здійснених у процесі прийняття цих рішень. Тобто пояснення відображають причинно-наслідкові між діями процесу прийняття рішень в системі штучного інтелекту. Тому використання пояснень забезпечує можливість успішного застосування отриманих у інтелектуальній інформаційній системі результатів при вирішенні практичних задач користувачів [4, 5]. Для побудови пояснень формується спрощена модель процесу прийняття рішення, яка відображає ключові, узагальнені залежності між даними, діями процесу і отриманим результатом. При побудові такої моделі залежності узагальнюються на заданому рівні деталізації, що потребує визначення схожих вхідних і проміжних даних та подальшого їх об'єднання в рамках класів еквівалентності. Клас еквівалентності даних містить схожі елементи, що мають однакові властивості. Ці дані можуть бути взаємозамінними у визначеному контексті. Тому використання класів еквівалентності дає можливість компактно описати групу даних зі схожими властивостями та узагальнити пов'язані із ними причинно наслідкові залежності.

Аналіз останніх досліджень і публікацій. Сучасні дослідження з розробки пояснень в системах штучного інтелекту були представлені в програмі зрозумілого штучного інтелекту (XAI) від агентства DARPA [6, 7]. Існуючі підходи до побудови пояснень пов'язані із визначенням важливості особливостей вхідних даних [8], розподілом внеску властивостей системи штучного інтелекту у кінцеве рішення [9], побудовою спрощеної локальної моделі прийняття рішення на основі виявлення впливу відхилень у вхідних даних на кінцевий результат[10]. Однак існуючі підходи мають ряд недоліків, пов'язаних із локалізацією пояснення, акцентуванням на пояснення в першу чергу результату відносно вхідних даних [8], а не процесу його формування, а також обчислювальною складністю побудови пояснення [9]. Для подолання цих недоліків доцільно розробити підхід, який би єдиним способом, на основі каузальних зв'язків [11, 12] описував би залежності між вхідними даними і результатом, а також між діями процесу прийняття рішення. Таким чином, актуальною є задача побудови причинно-наслідкових залежностей між класами

еквівалентності даних інтелектуальної системи, оскільки її вирішення дає можливість побудувати пояснення щодо причин дій процесу прийняття рішення.

Мета та задачі дослідження. Мета роботи полягає у розробці підходу до побудови можливісної каузальної моделі пояснення з урахуванням класів еквівалентності вхідних, проміжних і результуючих даних. Вирішення даної задачі створює умови для побудови пояснень у формі причинно-наслідкових залежностей на основі доступної інформації про властивості вхідних даних, а також про властивості отриманих в системі штучного інтелекту результатів.

Для досягнення мети вирішуються такі задачі: розробка моделі каузальної залежності між класами еквівалентності вхідних та вихідних даних; розробка методів побудови класів еквівалентності даних процесу прийняття рішення та побудови причинно-наслідкового представлення пояснення.

Модель каузальної залежності між класами еквівалентності даних

Клас еквівалентності представляє собою підмножину елементів, які є взаємозамінними (або однаковими) згідно відношенню еквівалентності. Використання класів еквівалентності дає можливість ефективно об’єднати схожі дані, що представляють певний об’єкт, представляючи кожен елемент даних як представника відповідного класу, що відображає цей об’єкт. Це дає можливість групувати та структурувати інформацію. В результатів при побудові каузальних залежностей для формування пояснень можна розглядати лише представників кожного класу еквівалентності замість перебору всіх можливих комбінацій значень вхідних даних.

При побудові класів еквівалентності для формування пояснень доцільно враховувати структуру вхідних даних та отриманих в системі штучного інтелекту рішень. Узагальнена структура вхідних даних представлена на рис. 1.

Дана вимога обумовлюється такими рівнями знань користувача щодо процесу прийняття рішень та безпосередньо рішення системи штучного інтелекту: знання про особливості використання рішення системи; знання про загальні принципи роботи інтелектуальної системи.

Перша група охоплює знання щодо умов та обмежень використання рішення системи. Обмеження задають допустиму сферу використання рішення, а умови визначають найбільш ефективні способи використання.

Друга група охоплює базові знання про процес прийняття рішення:

– основні закономірності, які лежать в основі процесу прийняття рішення; ці закономірності задають обмеження на допустимі пояснення;

– основні кроки процесу прийняття рішення; ці кроки визначають ключові причинно-наслідкові залежності, які обумовлюють властивості рішення.

На рис.1 показано, що класи еквівалентності для вхідних та вихідних об’єктів визначаються як з урахуванням значень змінних, так і з урахуванням відношень між властивостями об’єкта, представленого відповідними змінними, та безпосередньо об’єктом.

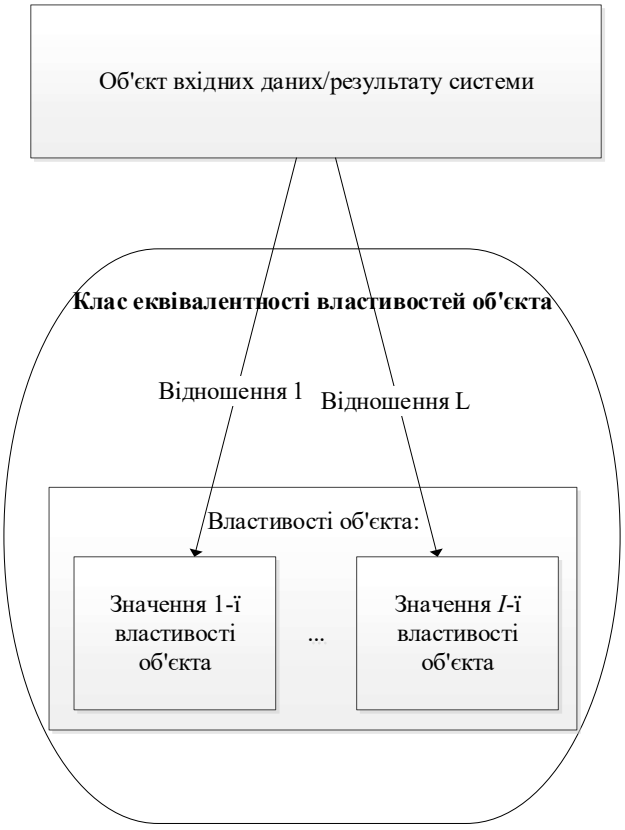

Рис. 1. Узагальнена структура вхідних та вихідних даних процесу прийняття рішення в системі штучного інтелекту

Тобто при визначенні еквівалентних властивостей об’єкта їх можна об’єднати з урахування їх відношення до об’єкта в цілому. Наприклад, для об’єкта «процесор», властивості якого використовуються в якості вхідних даних в рекомендаційній системі, можна визначити два відношення уточнення (рис. 2).

В цілому даний підхід дає можливість побудувати структуру класів еквівалентності, що відображає особливості як вхідних об’єктів, так і рішення системи штучного інтелекту, що дає можливість визначити причинно-наслідкові залежності на потрібному рівні деталізації та зробити прозорими для користувача цільові аспекти процесу прийняття рішення.

Тобто згідно запропонованого підходу класи еквівалентності враховують структуру об’єкта. Властивості об’єкта будуть еквівалентними у тому випадку, якщо вони мають однакове або мають схоже значення та однакове відношення до об’єкта в цілому. Відповідно, структура вхідного об’єкта може бути представлена в організаційному аспекті (зазвичай як ієрархія) і у процесному аспекті (зазвичай як виділена послідовність ознак).

Запропонований перелік відношень для структуризації вхідних даних та результату роботи інтелектуальної системи наведено в табл. 1 та 2.

Відношення першої групи дають можливість описати ієрархічну структуру вхідних, проміжних та

результуючих даних при побудові пояснення в системі штучного інтелекту. Такий підхід дає можливість узагальнити дані на рівні відповідного класу еквівалентності при формуванні пояснення і, тим самим, надати пояснення без непотрібної деталізації, у відповідності до рівня сприйняття користувача.

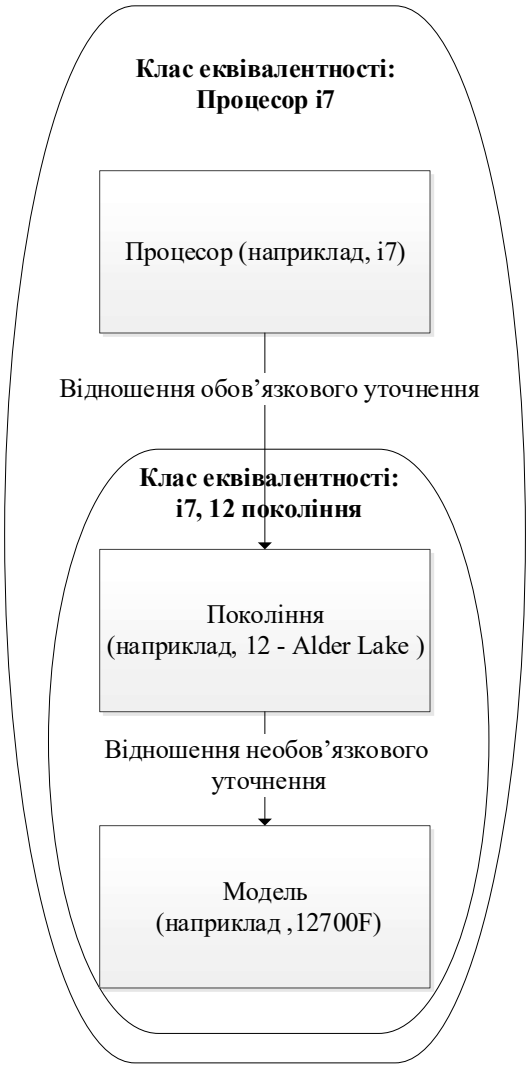

Рис. 2. Приклад відношень уточнення для класів еквівалентності

Відношення другої групи дають можливість відібрати (або виключити) конкретні значення властивостей при побудові каузальних правил для пояснення в системі штучного інтелекту. На відміну від першої групи, де виконується узагальнення причин рішення відповідно до вибраного ієрархічного рівня опису об'єктів, щодо яких приймається рішення в інтелектуальній системі, друга група відношень дає можливість визначити ключові ознаки – причини рішення, або ж виключити ознаки рішення, які не є суттєвими для його практичного використання.

Позначимо через l значення $k-$ властивості об'єкта, що використовується у процесі прийняття рішення, а через r_j^i зв'язок між $i-$ причиною та $j-$ рішенням (проміжним або кінцевим).

Таблиця 1 – Ієрархічні відношення в класах еквівалентності для опису вхідних, проміжних і вихідних даних при побудові пояснень в системі штучного інтелекту

Відношення	Опис
1. Відношення необов'язкового уточнення	1) Властивість є необов'язковою ознакою об'єкта. 2) Якщо значення властивості є причиною рішення, то і об'єкт в цілому є причиною рішення в інтелектуальній системі. 3) Якщо об'єкт у поточному стані є причиною рішення, то значення властивості може не бути такою причиною. 4) Приклад для рекомендаційної системи: якщо поточна модель процесора i7-12700F обумовлює рекомендацію моделі ноутбука користувачеві, то наявність процесора i7 є причиною рекомендації ноутбука. Проте зворотна залежність не є обов'язковою: якщо ноутбук вибрано тому, що він має процесор i7, то це не означає, що причиною є конкретна модель 12700F.
2. Відношення обов'язкового уточнення	1) Властивість є обов'язковою ознакою об'єкта. 2) Якщо поточне значення властивості є причиною рішення, то це однозначно свідчить, що об'єкт в цілому є причиною рішення в інтелектуальній системі. 3) Якщо об'єкт у поточному стані є причиною рішення, то значення властивості також є такою причиною. 4) Приклад для рекомендаційної системи: якщо розмір екрану в 14" є однією з причин рекомендації ноутбуку, то і ноутбук з екраном в 14 дюймів є умовою рекомендації. Аналогічно, зворотне також буде вірним.

Тоді клас еквівалентності $[f_i]$ на основі відношення необов'язкового уточнення формується таким чином:

$$[f_i] = \left\{ f_i^{k,l} : \exists r_j^i \right\}. \quad (1)$$

Клас еквівалентності на основі відношення обов'язкового уточнення формується таким чином:

$$[f_i] = \left\{ f_i^{k,l} : (\forall k \forall l) \exists r_j^i \right\}. \quad (2)$$

Тобто в першому випадку не всі елементи $f_i^{k,l}$ класу еквівалентності можуть бути використані в якості причини для проміжного або фінального рішення системи штучного інтелекту. В другому випадку всі значення властивостей можуть розглядатись як суттєві причини.

Слід зазначити, що в рамках можливісного підходу відношення обов'язкового і необов'язкового уточнення можуть бути визначені за результатами обчислення можливості та необхідності значень властивостей для визначення причин отриманого рішення.

Тобто запропонований підхід дозволяє в офлайн-режимі сформувати набір каузальних правил, що є елементами пояснення, та в подальшому використати ці правила в онлайн-режимі вже безпосередньо для побудови пояснень.

Таблиця 2 – Логічні відношення в класах еквівалентності для опису вхідних, проміжних і вихідних даних при побудові пояснень в системі штучного інтелекту

Відношення	Опис
1. Відношення вимоги	1) Властивість є обов'язковою ознакою для побудови каузальної залежності у складі пояснення. 2) Зазвичай дане відношення вимагає деталізації опису об'єкта у вигляді конкретного значення властивості. 3) Приклад для рекомендаційної системи: вимога визначити конкретне значення розміру SSD - пам'яті.
3. Відношення виключення	1) Властивість не є необхідною при побудові пояснень. 2) Відношення дає можливість виключити несуттєві властивості з каузального правила, що визначає пояснення. 3) Приклад для рекомендаційної системи: наявність тач-скріну не є необхідною ознакою при рекомендації ноутбука.
3. Відношення кон'юнкції	1) Набір конкретних значень властивостей, які визначають причини рішення інтелектуальної системи. 2) Відношення дає можливість визначити набір причин рішення на різних рівнях деталізації даних..

Логічні відношення при побудові класу еквівалентності використовуються наступним чином. Клас еквівалентності на основі відношення вимоги має вигляд:

$$\left[f_i\right]=\left\{f_i^{k,l}:(\forall i)\exists r_j^i\right\}. \quad (3)$$

Клас еквівалентності на основі відношення виключення має вигляд:

$$\left[f_i\right]=\left\{f_i^{k,l}:(\forall i)\neg\exists r_j^i\right\}. \quad (4)$$

Тобто згідно (3) кожне із значень властивостей використовується при побудові пояснень.

В класі еквівалентності (4) значення властивостей не враховуються ні в одному з правил r_j^i.

Відношення кон'юнкції використовується для класу еквівалентності, елементи якого об'єднують декілька значень властивості як єдину умову для каузального правила у складі пояснення:

$$\left[f_i\right]=\left\{\left\{f_i^{k,l}\wedge\ldots\wedge f_i^{K,L}\right\}\middle|k=\overline{1,K},\ l=\overline{1,L}\ (\forall i)\exists r_j^i\right\}. \quad (5)$$

Слід зазначити, що $f_i^{k,l}$ у виразі (5) можуть належати іншим класам еквівалентності, що створює умови для структурованого представлення вхідних, проміжних та результуючих даних на необхідному для побудови пояснення рівні деталізації/узагальнення.

Вирази (1) – (5) визначають класи еквівалентності при формуванні вхідних класів еквівалентності. Вихідні класи еквівалентності $\left[f_j\right]$, в тому числі фінальне рішення інтелектуальної системи можуть бути представлені на основі переважно відношення (5).

Тоді каузальні залежності, що входять до складу пояснення, можуть бути представлені через причинно-наслідкових зв'язок між класами еквівалентності у такому узагальненому вигляді:

$$\left[f_i\right]r_j^i\left[f_j\right]. \quad (6)$$

З урахуванням кон'юнкції елементів та класів еквівалентності (5), представлення (6) відображає причинно-наслідковий зв'язок між структурованим за рівнями ієрархії та за важливістю набором причин $\left[f_i\right]$ та аналогічно структурованим результатом $\left[f_j\right]$.

Можливісний причинно-наслідковий зв'язок визначається, згідно теорії можливостей, з урахуванням можливості Π_j^i та за умови необхідності N_j^i даного правила. Тому можливісний зв'язок між класами еквівалентності має вигляд:

$$\left[f_i\right]\Pi_j^i\left[f_j\right]\middle|N_j^i. \quad (7)$$

Згідно (7), каузальна залежність між класами еквівалентності є можливою з оцінкою можливості Π_j^i за умови заданого рівня необхідності N_j^i.

Методи побудови можливісно-каузальних залежностей для пояснення на основі класів еквівалентності даних.

Запропонований підхід до побудови можливісної каузальної моделі пояснення з урахуванням класів еквівалентності вхідних, проміжних і результуючих даних імплементовано у вигляді методу побудови класів еквівалентності даних процесу прийняття рішення та методу формування причинно-наслідкового представлення пояснення.

Перший метод формує класи еквівалентності згідно умов, представлених виразами (1) – (5). Метод складається з наступних етапів,

Етап 1. Визначення відношень між даними згідно табл. 1.

Етап 2. Визначення класів еквівалентності для вхідних даних.

Етап 3. Визначення класів еквівалентності для вихідних даних.

Етап 4. Побудова множини темпоральних правил F- та X-типу для визначених класів еквівалентності.

Правила F-типу використовуються для відображення зв'язку між вхідними даними та результатом інтелектуальної системи, а правила X – лише для зв'язку у часі між проміжними даними процесу прийняття рішення.

Результатом методу є пари класів еквівалентності, упорядковані у часі у відповідності до процесу прийняття рішень у системі штучного інтелекту.

Метод формування можливісно-каузальних правил містить такі етапи.

Етап 1. Формування каузальних правил на базі темпоральних правил F-типу та X-типу.

Етап 2. Розрахунок можливості причинно-наслідкового зв'язку для правил.

Етап 3. Розрахунок необхідності каузальних правил.

Етап 4. Відбір правил за пороговим значенням необхідності.

Етап 5. Упорядкування правил за значенням можливості.

Результатом методу є можливісні каузальні правила, що відображають можливість зв'язків між класами еквівалентності вхідних, проміжних та результуючих даних.

Розглянемо приклад реалізації першого методу для побудови пояснень щодо рекомендації ноутбуків в системі електронної комерції.

Вхідні дані: ноутбуки на процесорах $i3-i9$ з кінцевими множинами значень об'єму оперативної пам'яті, жорсткого диску, розміру та роздільної здатності екрану, тощо. Кожна з моделей, крім наведених значень змінних, містить інформацію про конкретну модель процесора.

На етапі 1 формуються відношення між значеннями змінних. Зокрема, відношення обов'язкового уточнення виду $Процесор = \{i3, i5, i7, i9\}$, $i7 = \{i711, i712, i713\}$, відношення необов'язкового уточнення виду $i711 = \{i7-1185, i7-1165, i7-1155, ...\}$, відношення вимоги виду $Екран = \{13", 14", ...\}$, тощо. Ці відношення формуються з урахуванням результату роботи системи: рекомендованої моделі комп'ютера. Така модель містить відповідні параметри процесора, розміру екрану, пам'яті, тощо. Обов'язкові вхідні дані (відношення включення) вибираються з урахуванням дій користувача при пошуку ноутбука, тобто з урахуванням фільтру, який поставив користувач. Наприклад, фільтру по розміру екрану. Отримані відношення визначають класи еквівалентності вхідних даних. Для вихідних даних класи еквівалентності формуються аналогічно, з урахуванням всіх параметрів рекомендованого ноутбука. На етапі 4 формуються темпоральні залежності виду:

$$\begin{pmatrix} Клас\ процесора: \\ i7 \end{pmatrix} F \begin{pmatrix} Клас\ результату: \\ модель\ ноутбука \end{pmatrix}. \quad (8)$$

Ця залежність відображає послідовність дій в часі: спочатку вибір класу об'єктів, що відображають властивості ноутбука, а потім клас рекомендованого результату. В даному випадку маємо залежність F-типу, оскільки проміжні дані не розглядаються.

На другій фазі для кожного темпорального правила розраховується значення можливості і необхідності згідно теорії можливості.

Можливість, наприклад, розраховується через найбільшу ймовірність продажу ноутбуку з процесором відповідного класу ($i7, i9$ тощо). Необхідність розраховується на основі ймовірності продажу моделей з альтернативними процесорами. Вказані ймовірності розраховуються на основі даних про продажі системи електронної комерції.

Результатом другого методу є множина залежностей виду (клас еквівалентності для властивості процесора – рекомендована модель).

В подальшому наведені залежності можуть бути об'єднані згідно (5). Відповідно зміняться значення ймовірностей і другий метод необхідно буде виконати повторно.

Висновки. Запропоновано можливісну модель каузальної залежності, що містить причинно-наслідковий зв'язок між класами еквівалентності вхідних або проміжних та результуючих даних, отриманих у процесі прийняття рішення в системі штучного інтелекту. Даний причинно-наслідковий зв'язок базується на темпоральній залежності між даними та враховує оцінки можливості і необхідності такої залежності. Модель створює умови для пояснення можливих причин отриманого рішення на основі узагальнення як вхідних, так і проміжних даних процесу прийняття рішення в інтелектуальній системі.

Запропоновано комплекс методів для побудови класів еквівалентності даних процесу прийняття рішення та побудови причинно-наслідкового представлення пояснення, що встановлює каузальний зв'язок між класами еквівалентності. У процесі побудови класів еквівалентності встановлюються відношення уточнення даних, вимоги або виключення даних, а також кон'юнкції даних. У процесі побудови причинно-наслідкового представлення пояснення розраховується можливість та необхідність такої залежності. В практичному плані розроблений комплекс методів дає можливість побудувати пояснення на основі доступної інформації про отримані рішення та вхідні й проміжні дані, які були використані ля формування цих рішень.

Список використаної літератури

1. Engelbrecht Andries P. *Computational Intelligence: An Introduction*. NJ: John Wiley & Sons, 2007. 632 p.
2. Alonso J.M., Castiello C., Mencar C. A Bibliometric Analysis of the Explainable Artificial Intelligence Research Field. In: Medina, J., et al. Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations. *IPMU. Communications in Computer and Information Science*. 2018. Vol. 853. P. 3–15.
3. Чалий С. Ф., Лещинська І. О. Концептуальна ментальна модель пояснення в системі штучного інтелекту. *Вісник Національного технічного університету «ХПІ». Серія: Системний аналіз, управління та інформаційні технології.* Харків: НТУ «ХПІ», 2023. № 1 (9). С. 70–75.
4. Tintarev N., Masthoff J. A survey of explanations in recommender systems. The 3rd international workshop on web personalisation, recommender systems and intelligent user interfaces (WPRSIUI'07). 2007. P. 801-810.
5. Camburu O.M, Giunchiglia E., Foerster J., Lukasiewicz T., Blunsom P. Can I trust the explainer? Verifying post-hoc explanatory methods. 2019. *arXiv*:1910.02065.
6. Gunning D., Aha D. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*. 2019. Vol. 40(2). P. 44-58.
7. D. Gunning, E. Vorm, J. Wang, M. Turek, "Darpa's Explainableai(XAI) Program: a Retrospective", *Applied AI Letters.* Vol. 2, no. 4, 2021. https://doi.org/10.1002/ail2.61
8. Friedman J.H. Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics. 2001. Vol. 29(5), P.1189-1232.

9. Lundberg S.M., Lee, S.I. A Unified Approach to Interpreting Model Predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017).* 2017. P. 4765-4774.
10. Ribeiro M.T., Singh S. Guestrin, C. Why Should I Trust You?: Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 2016. P. 1135-1144.
11. Chalyi, S., & Leshchynskyi, V. Temporal-oriented model of causal relationship for constructing explanations for decision-making process. *Advanced Information Systems.* 2022 №6(3), P. 60–65.
12. Chalyi S, Leshchynskyi V. Probabilistic counterfactual causal model for a single input variable in explainability task. *Advanced Information Systems.* 2022. №7(3), P.54–59. https://doi.org/10.20998/2522-9052.2023.3.08

References (transliterated)

1. Engelbrecht Andries P. *Computational Intelligence: An Introduction*. NJ: John Wiley & Sons, 2007. 632 p.
2. Alonso J.M., Castiello C., Mencar C. A Bibliometric Analysis of the Explainable Artificial Intelligence Research Field. In: Medina, J., et al. Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations. *IPMU. Communications in Computer and Information Science*. 2018, vol. 853. pp. 3–15.
3. Chalyi S. F., Leshchynska I. O. Kontseptualna mentalna model poiasnennia v systemi shtuchnoho intelektu. *Visnyk Natsionalnoho tekhnichnoho universytetu «KhPI». Seriia: Systemnyi analiz, upravlinnia ta informatsiini tekhnolohii* [Bulletin of the National Technical University "KPI". Series: System Analysis, Control and Information Technology]. Kharkiv, NTU "KhPI" Publ., 2023, no. 1 (9), pp. 70–75.
4. Tintarev N., Masthoff J. A survey of explanations in recommender systems. The 3rd international workshop on web personalisation, recommender systems and intelligent user interfaces (WPRSIUI'07). 2007, pp. 801-810.
5. Camburu O.M, Giunchiglia E., Foerster J., Lukasiewicz T., Blunsom P. Can I trust the explainer? Verifying post-hoc explanatory methods. 2019. *arXiv*:1910.02065.
6. Gunning D., Aha D. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine.* 2019, vol. 40(2). pp. 44-58.
7. D. Gunning, E. Vorm, J. Wang, M. Turek, "Darpa's Explainableai(XAI) Program: a Retrospective", *Applied AI Letters*. 2021, vol. 2, no. 4, https://doi.org/10.1002/ail2.61
8. Friedman J.H. Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics. 2001, vol. 29(5), pp.1189-1232.
9. Lundberg S.M., Lee, S.I. A Unified Approach to Interpreting Model Predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017).* 2017, pp. 4765-4774.
10. Ribeiro M.T., Singh S. Guestrin, C. Why Should I Trust You?: Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD),* 2016, pp. 1135-1144.
11. Chalyi, S., & Leshchynskyi, V. Temporal-oriented model of causal relationship for constructing explanations for decision-making process. *Advanced Information Systems.* 2022 №6(3), P. 60–65.
12. Chalyi S, Leshchynskyi V. Probabilistic counterfactual causal model for a single input variable in explainability task. *Advanced Information Systems.* 2022, no. 7(3), pp.54–59. https://doi.org/10.20998/2522-9052.2023.3.08

Надійшла (received) 10.05.2024

UDC 004.8:004.9

S. F. CHALYI, Doctor of Technical Sciences, Full Professor, Kharkiv National University of Radio Electronics, Professor of the Department of Information Control System, Kharkiv; e mail: serhii.chalyi@nure.ua, ORCID: https://orcid.org/0000-0002-9982-9091

V. O. LESHCHYNSKYI, Candidate of Technical Sciences (PhD), Associate Professor, Kharkiv National University of Radio Electronics, Associate Professor at the Department of Software Engineering, Kharkiv; e–mail: volodymyr.leshchynskyi@nure.ua, ORCID: https://orcid.org/0000-0002-8690-5702

CONSTRUCTION OF PROBABILISTIC CAUSAL RELATIONSHIPS BETWEEN EQUIVALENCE CLASSES OF DATA IN AN INTELLIGENT INFORMATION SYSTEM

The subject of this research is the processes involved in generating explanations for decision-making in artificial intelligence systems. Explanations in such systems enable the decision-making process to be transparent and comprehensible for the user, thereby increasing user trust in the obtained results. The aim of this work is to develop an approach for constructing a probabilistic causal explanation model that takes into account the equivalence classes of input, intermediate, and resulting data. Solving this problem creates conditions for building explanations in the form of causal relationships based on the available information about the properties of input data as well as the properties of the results obtained in the artificial intelligence system. To achieve this aim, the following tasks are addressed: developing a causal dependency model between the equivalence classes of input and output data; developing methods for constructing equivalence classes of data in the decision-making process and a method for constructing causal explanations. A probabilistic model of causal dependency is proposed, which includes a causal relationship between the equivalence classes of input or intermediate and resulting data obtained during the decision-making process in the artificial intelligence system. This relationship considers the estimates of the possibility and necessity of such a dependency. The model creates conditions for explaining the possible causes of the obtained decision. A set of methods for constructing equivalence classes of data in the decision-making process and for constructing causal explanations is proposed, establishing a causal relationship between the equivalence classes. When constructing equivalence classes, relations of mandatory and optional data refinement, requirements or exclusions of data, as well as data conjunctions, are established. When constructing causal explanations, the possibility and limitations of the necessity of such a dependency are calculated, allowing explanations to be built based on the available information about the obtained decisions and the input and intermediate data used to form these decisions.

Keywords: Causal dependency, cause-and-effect relationship, temporal dependency, possibility, necessity, explanation, artificial intelligence system, intelligent system, explainable artificial intelligence, information system.

Повні імена авторів / Author's full names

Автор 1 / Author 1: Чалий Сергій Федорович**,** Chalyi Serhii Fedorovych
Автор 2 / Author 2: Лещинський Володимир Олександрович, Leshchynskyi Volodymyr Oleksandrovich