https://doi.org/10.31649/1997-9266-2019-143-2-78-90

УДК 004.94

Т. Д. Польгуль[1]
А. А. Яровий[1]

АНАЛІЗ РІЗНОРІДНИХ ДАНИХ В ІНТЕЛЕКТУАЛЬНИХ СИСТЕМАХ ВИЯВЛЕННЯ ШАХРАЙСТВА

[1]Вінницький національний технічний університет

Авторами статті шахрайство розглядається як аномалія в даних. Розроблено метод аналізу різнорідних даних в інтелектуальних системах виявлення шахрайства. Формалізовано процес виявлення шахрайства як аномалії в різнорідних даних при інсталюванні мобільних додатків з використанням теорії множин, що дозволило здійснити подальший аналіз даних у таких системах. Запропоновано математичну модель процесу аналізу різнорідних даних, алгоритм аналізу різнорідних даних, метод аналізу різнорідних даних на основі запропонованих шкал та коефіцієнтів, що дозволили обробляти різноформатні вхідні дані — різних метрик, шаблонів, розмірності, що у процесі аналізу дає можливість сформувати узагальнений шаблон шахрая. Розроблений метод використовує бази даних та бази знань, завдяки яким формується узагальнений шаблон шахрая, наявність якого дозволяє прискорити виявлення шахраїв у нових наборах даних та виявляти навіть неявних шахраїв. Запропонований метод розроблений з метою його використання в інтелектуальних системах виявлення шахрайства на основі аномалії в даних, які, на відміну від існуючих, дозволять здійснити аналіз різнорідних даних, на основі яких приймаються рішення про шахрайство, зменшити розмірності даних та провести класифікацію користувачів. Проведено експериментальні дослідження запропонованого методу аналізу різнорідних даних у межах виявлення шахрайства як аномалії в різнорідних даних та класифікаційної моделі, розробленої з використанням повністю зв'язаних глибоких нейронних мереж з трьома прихованими шарами з використанням розробленого програмного забезпечення та з використанням репрезентативної вибірки. Запропоновано схему експериментального дослідження виявлення аномалій в різнорідних даних при інсталюванні мобільних додатків, в основі якої є представлений метод аналізу різнорідних даних. Показано ефективність використання запропонованого методу у системі виявлення шахрайства, точність класифікації якої склала 99,14 %, точність виявлення шахраїв — 82,76 %. Проте зі збільшенням правил у розробленій базі знань, що буде відбуватися з кожним запуском на нових даних, збільшуватиметься точність системи.

Ключові слова: виявлення шахрайства, виявлення аномалій, модель класифікації, метод аналізу різнорідних даних.

Вступ

Для сучасного ІТ-ринку характерно просування компаніями-розробниками мобільних додатків. За таку процедуру компанія повинна витрачати достатньо великі кошти на маркетингові кампанії. Для оцінки дієвості маркетингових кампаній зазвичай використовують перевірку приведеної нею кількості інсталювань мобільного додатку. Проте, слід зазначити, що частина або вся множина інсталювань мобільних додатків могла бути здійснена шахрайським способом. При цьому компанія-розробник втратить кошти, витрачені на таку «шахрайську» маркетингову кампанію та не отримає органічних користувачів, заради чого і проводяться такі кампанії. Тому й виникла необхідність побудови автоматизованої інтелектуальної системи виявлення шахрайства при інсталюванні мобільних додатків, яка б дозволила визначати шахраїв як з відомими, так і з новими шаблонами, які непомітні чи невідомі експертам. Для побудови такої системи виникла потреба розробки методу аналізу різнорідних даних з метою надання системі інтелектуальної складової, а саме — її здатності до подальшого самонавчання і можливості адаптуватися.

Постановка задачі

Актуальною в цьому напрямку є задача розробки системи для автоматичного виявлення маркетингових кампаній та користувачів, які використовують шахрайські способи інсталювання. Ще однією актуальною проблемою є те, що існуючі системи виявлення шахрайства не є адаптивними

до появи нових видів шахрайства. Проте цю проблему можна вирішити, навчившись інтелектуально аналізувати різнорідні дані. Адже всі наявні дані в таких системах є різнорідними, а саме різних форматів, шаблонів та розмірності. Для цього актуальною є розробка методу аналізу різнорідних даних в інтелектуальних системах виявлення шахрайства.

У разі розробки таких систем та зокрема методу аналізу різнорідних даних у таких системах важливим є питання як визначити шахрайство. Розглядатимемо шахрайство як аномалію у вхідних даних, а саме — як навмисне породження аномалії в даних про досліджуваний процес сторонньою особою (шахраєм) або механізмом з певною метою [1]. А поняття аномалії, у свою чергу, присутнє та досліджене практично у всіх областях різними науковими школами. Зазвичай у сучасній науці і техніці аномалія розглядається як шаблон даних, який не відповідає визначеному поняттю «нормальної» поведінки (заданому шаблону) досліджуваного процесу — так визначають вчені з університету Мінесотта. Джонсон (Johnson) визначає аномалію як спостереження у наборі даних, яке суперечить іншій частині цього набору даних, а Хокінз (Hawkins) у [2] визначає аномалію (викид) як спостереження, яке відхиляється від інших спостережень настільки, що виникають підозри, що він був породжений іншим механізмом.

Тому в роботі розроблено метод аналізу різнорідних даних у межах створення інтелектуальної системи виявлення шахрайства як аномалії в даних при інсталюванні мобільних додатків.

Мета дослідження

Метою роботи є розробка методу аналізу різнорідних даних з метою подальшого його використання для розробки інтелектуальної автоматизованої системи ефективного виявлення шахрайства як аномалій в різнорідних даних при інсталюванні мобільних додатків.

Для досягнення поставленої мети необхідно вирішити такі задачі:

– проаналізувати існуючі підходи навмисного внесення аномалій в дані (шахрайства) з метою подальшої розробки методу виявлення аномалій в різнорідних даних з їх урахуванням;

– формалізувати процес виявлення шахрайства як аномалії в даних при інсталюванні мобільних додатків;

– виявлення повної інформації про користувача;

– аналіз всіх наявних різнорідних даних різних шаблонів, розмірності, метрик з використанням методу аналізу різнорідних даних, розробка якого є необхідною;

– розробити метод аналізу різнорідних даних;

– оцінити точність виявлення аномалій в різнорідних даних у заданій репрезентативній вибірці користувачів з використанням методу аналізу різнорідних даних.

Аналіз існуючих підходів виявлення шахрайства як аномалій в даних

На сьогодні відомі такі шахрайські види інсталювання мобільних додатків [3]—[6]:

– кліковий спам (click spamming) — генерація підроблених запитів кліків програмним способом;

– мобільне викрадення (mobile hijacking) — виконання несанкціонованих дій справжнім мобільним додатком, який встановлений на пристрої органічного користувача. Так наприклад, можливе формування кліків від імені користувача та запуск прихованих оголошень додатком у фоновому режимі. Крім того, програма працюватиме за сценарієм, що максимально імітує поведінку органічного користувача;

– ферми дій (action farms). Шахраї винагороджують людей по всьому світу за установлення мобільних додатків у ручному режимі, тобто фактично відбувається найняття людей для того, щоб вони встановлювали мобільні додатки.

Фахівці з програмного забезпечення часто використовують популярну платформу Kochava [7], яка виявляє шахрайство при інсталюванні мобільних додатків за власною методикою. А саме, у платформі визначені певні правила, за якими відсікаються шахраї. Критерії відбору шахраїв такі: великі обсяги кліків з однієї IP-адреси; велика кількість кліків з одного пристрою; відхилення середнього часу інсталювання (mean-time-to-install outliers — MTTI); відхилення часу інсталювання (Time-to-install Outliers – TTI); географічні кліки/дельта при встановленні; платформа кліку/невідповідність інсталяцій; відповідність множинних хеш-атрибутів; кліки, що складаються з реклами; анонімні інсталювання; надходження непідтверджених інсталювань; надходження непідтверджених покупок.

Але такий підхід дозволяє визначати лише певні шаблони шахрайства, насправді ж кожного дня шахраї придумують нові способи інсталювання мобільних додатків, які також необхідно ав-

томатизовано виявляти. Для того, щоб розширити можливості, які пропонує Kochava, можна розглянути інші методи та системи-аналоги, більшість з яких детальніше розглянуто у роботі [5]. На погляд авторів найвдалішими для виявлення аномалій у великих масивах даних є методи машинного навчання для обробки Big Data, оскільки вхідні дані — це великі масиви даних. Тому розглянемо відомі методи машинного навчання для обробки Big Data для виявлення аномалій [1], [8]—[10], які можна розділити на такі групи:

– методи класифікації. Для задачі, що розглядається, найактуальнішими є експертні системи [11], методи класифікації на основі нейронних мереж, метод k-найближчих сусідів;

– методи кластеризації [13], [14]. У багатьох роботах ці методи поділяють на ієрархічні (таксономія) та неієрархічні або ж на чіткі та нечіткі. Найбільше використовуються k-means clustering, графові методи, алгомеративна ієрархічна кластеризація;

– статистичні методи. Серед цих методів можна виділити спектральний метод, непараметричне статистичне моделювання та параметричне статистичне моделювання [15], [16].

Для побудови інтелектуальної системи виявлення шахрайства при інсталюванні мобільних додатків у роботі доцільно обрати методи класифікації, та зокрема експертні системи, оскільки вони використовуються у всіх подібних областях. Слід зазначити, що більшість методів класифікації очікують отримати на вхід однорідні дані. Тому розробка методу аналізу різнорідних даних є доцільною, оскільки дозволить привести всі дані до однорідних, а також використати інформацію про усі наявні дані. Такий підхід усуне можливе упущення специфічних характеристик шахраїв, що допоможе побудувати точнішу систему виявлення шахраїв навіть за появи нових видів шахрайства.

Формалізація процесу виявлення аномалій в даних як шахрайства при інсталюванні мобільних додатків

Пошук аномалій в даних оснований на аналізі всієї множини даних про маркетингові кампанії, користувачів та послідовності їх дій, технічні та часові характеристики, які є важливими при інсталюванні мобільних додатків. Тобто ці дані є різнорідними. Під різнорідністю в більшості розглянутих методів розумітимемо дані різних типів (числові, категоріальні, бінарні) та розмірності, які не можливо рівноцінно порівняти між собою, та які приймають значення різних діапазонів. Відомо, що на сьогодні у цій області дослідження існує два підходи до вирішення проблеми різнорідності даних. Перший підхід — це перехід від різнорідних до однорідних даних з використанням сучасних методик, таких як багатовимірне шкалювання, one-hot encoding, розглянутих у роботі [1]. Проте використання багатовимірного шкалювання втрачає точність вхідних даних. У свою чергу використання методу one-hot encoding для інтерпретації якісної ознаки у вигляді кількісної потребує знання усіх можливих категорій цієї ознаки. Проте у випадку з такою якісною ознакою як, наприклад, IP-адреса не можливо завчасно знати всі можливі категорії. А також збереження матриці усіх можливих значень IP-адрес для кожного користувача використовуватиме надто багато ресурсів. Другий — це зменшення кількості даних. До методів другого типу можна віднести метод головних компонент (PCA — Principal Component Analysis). Зменшивши кількість вхідних важливих ознак з використанням методів другого типу, можна отримати неможливість обґрунтування прийнятого системою рішення на основі дійсних вхідних даних. Отже, вхідні дані для виявлення аномалій в даних є різнорідними. Сучасні методи для зняття різнорідності зменшують кількість даних, чим вносять невизначеність в них. Тому треба розробити метод, який зможе видобувати важливу інформацію з даних та не погіршити при цьому точність результатів, навіть якщо дані є різнорідними.

Таким чином, у разі виявлення аномалій в даних як шахрайства при інсталюванні мобільних додатків [3]—[6] важливим є розв'язання задачі аналізу всіх отриманих різнорідних даних, не втрачаючи при цьому точність розв'язання поставленої задачі. Авторами пропонується метод аналізу різнорідних даних, який дозволяє вирішити цю задачу.

Для розробки необхідного методу, формалізуємо процес виявлення аномалій в даних як шахрайства при інсталюванні мобільних додатків. Для початку формалізуємо поняття аномалії, для цього використаємо теорію множин. Аномалію будемо розглядати як групу (множину) даних Z, яка входить у множину вхідних даних A, задану множиною властивостей $P(a)$, та не відповідає заданим властивостям множини вхідних даних A. У задачі виявлення аномалій у вхідних наборах даних з мобільних додатків, аномальними будемо вважати ті елементи множини, які: не мають властивостей $P_1(x), P_2(x), ..., P_s(x)$, що визначають множину неаномальних даних X; не збігаються за властивостями групи даних; не збігаються за розмірністю та не входять в область гранично до-

пустимих значень — задамо це властивістю $P_p(a)$, яка має вигляд $(a \le \max\_val, a \ge \min\_val)$.

Формалізуємо вищесказане з використанням теорії множин для більшого розуміння. Вхідний набір даних можна подати математично у такому вигляді: $A = \{a \mid P_1(a), P_2(a), ..., P_s(a)\}$.

Розглянемо варіанти вхідних наборів даних систем прийняття рішень та систем штучного інтелекту, в яких наявні аномалії в даних:

1. Один з можливих варіантів, у якому чітко видно аномалію — це коли у вхідному наборі А є група елементів, які мають властивості набору А, не мають властивостей множини неаномальних даних Х, і всі у свою чергу мають різні властивості: $A = \{x_1, z_1, ..., x_n, ..., z_j, ... \mid x \in X, z \in Z\}$, де елементи підмножини $X = \{x_1, x_2, ..., x_n, ... \mid P_1(x), P_2(x), ..., P_s(x)\}$ мають властивості $P_1(x)$ і $P_2(x)$ і … і $P_s(x)$ та є не аномальними даними, елементи підмножини $Z = \{z_1, z_2, ..., z_j, ... \mid P_{a1}(z)$ або $P_{a2}(z)$ або...або $P_{j1}(z)\}$ у свою чергу не мають цих властивостей, що означає, що вони є аномальними у заданій множині даних А, та всі мають різні властивості — одну з: $P_{a1}(z)$ або $P_{a2}(z)$ або … або $P_{j1}(z)$, тому вони хаотично розкидані, відповідно $X \subseteq A$ та $Z \subseteq A$ та $X \cap Z = \varnothing$. Наочніше цей випадок показано у двовимірному просторі на рис. 1.

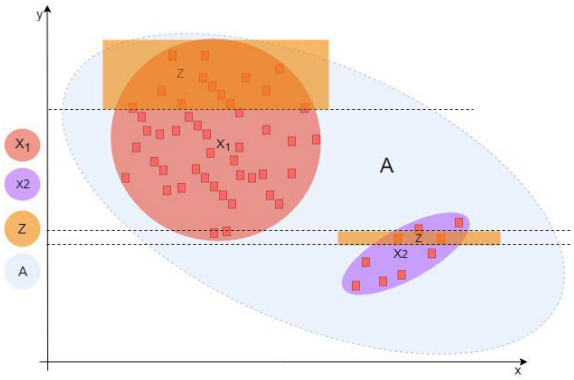

Рис. 1. Приклад аномалії типу 2 у двовимірному просторі: Х — підмножина неаномальних даних, сукупність підмножин $\{X_1, X_2\}$ є покриттям підмножини неаномальних даних Х; Z — підмножина аномальних даних

З рис. 1 видно, коли множина вхідних елементів має властивості, які обмежують елементи множини областю гранично допустимих значень. Так наприклад, властивості множини Х $P_1(x), P_2(x), ..., P_s(x)$ можуть мати такий вигляд:

$$P_1(x) = x \ge 100000;$$
$$P_2(x) = x \ne 205; \quad (1)$$
$$\ldots;$$
$$P_s(x) = x < 3001, x > 25.$$

2. Ще один з можливих варіантів, у якому чітко видно аномалію — це коли у вхідному наборі А є група елементів, які мають властивості набору А, але не мають властивостей набору неаномальних даних Х, проте всі мають одну спільну властивість $P_k(z)$, а отже, чітко виділяються на фоні інших даних: $A = \{x_1, x_2, ..., x_n, ..., z_1, z_2, ..., z_j, ... \mid x \in X, z \in Z\}$, де елементи підмножини $X = \{x_1, x_2, ..., x_n, ... \mid P_1(x), P_2(x), ..., P_s(x)\}$ мають властивості $P_1(x)$ і $P_2(x)$ і … і $P_s(x)$ та є не аномальними даними, елементи підмножини $Z = \{z_1, z_2, ..., z_j, ... \mid P_a(z)\}$ у свою чергу не мають цих властивостей, що означає, що вони є аномальними у заданій множині даних, та всі мають спільну властивість $P_a(z)$, завдяки якій вони виражено групуються в окремий клас, відповідно $X \subseteq A$ та $Z \subseteq A$ та $X \cap Z = \varnothing$.

3. Варіант, у якому спостерігається аномалія — це коли у вхідному наборі А є група елементів, які не збігаються за властивостями групи даних.

Для більшого розуміння цього варіанта показано діаграму Венна на рис. 2, де всі елементи підмножин X_1, X_2, X_3, X_4, X_5 є неаномальними, проте деякі об'єднання цих підмножин мають властивос-

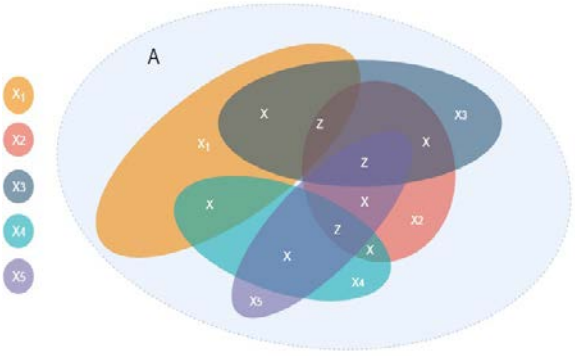

Рис. 2. Діаграма Вена, що зображає приклад аномалії типу 3: Х — підмножина неаномальних даних, сукупність підмножин $\{X_1, X_2, X_3, X_4, X_5\}$ є покриттям підмножини неаномальних даних Х; Z — підмножина аномальних даних

ті множини аномальних даних Z.

Формалізуємо цей вираз. Отже, А — множина вхідних даних, яка зображає цей випадок. Проте, для початку задамо множину неаномальних даних $X = \{x_1, x_2, ..., x_n, ... | P_1(x), P_2(x), ..., P_s(x)\}$, у свою чергу сукупність підмножин $\{X_1, X_2, ..., X_T\}$ є покриттям множини X.

Також задамо множину аномальних даних $Z = \{x_1 x_2 ... x_n ... | x_1, x_2, ..., x_n, ... \in X, P_{a1}(z), P_{a2}(z), ..., P_{aj}(z)\}$.

Тоді множину вхідних даних А можна задати як $A = \{x_1 x_2 ... x_n ... | x_1, x_2, ..., x_n, ... \in X, x_1 x_2 ... x_n ... \notin Z\}$.

4. Варіант, у якому спостерігається аномалія — це коли у вхідному наборі А є підмножина неаномальних даних X, яка є об'єднанням підмножин $X_1, X_2, ..., X_T$. Проте, деякі з підмножин X_i перевищують допустиму розмірність, або ж навпаки їх розмірність менша, ніж це можливо в неаномальних даних.

Математично такий випадок можна представити знову ж сукупністю підмножин $\{X_1, X_2, ..., X_T\}$, які є покриттям підмножини неаномальних даних X, що входить у множину вхідних даних А. У свою чергу деякі з підмножин $\{X_1, X_2, ..., X_T\}$ мають задані властивості розмірності, як наприклад $X_1 = \{x | count(x) < F\}$, де F — задане значення, $count(x)$ — кількість елементів у множині X_1.

Множина дорівнюватиме $A = \{a | a \in X\}$.

Розглянемо випадок шахрайства, як один з типів аномалії, який можна спостерігати при інсталюванні мобільних додатків. Коли у конкретній локації живе, наприклад, 300 000 жителів, а за один день з цієї локації пройшло 1 000 000 інсталювань. Тоді множина X_1, яка характеризуватиметься обмеженням $count(x) < 300000 + \varepsilon$ з одного пристрою за один день, вважатиметься аномальною, оскільки значно перевищуватиме задане обмеження розмірності.

Розглянувши приклади аномалій у вхідних наборах даних, формалізуємо вхідні дані системи виявлення шахрайства при інсталюванні мобільних додатків. Отже, матимемо множину користувачів $S = \{a_1, a_2, ..., a_v | a \in A\}$, в якій вхідні дані по кожному з користувачів задаються множиною $A = \{x_1, z_1, x_2, z_2, ..., x_n, ..., z_j, ... | x \in X, z \in Z\}$, де елементи підмножини $X = \{x_1, x_2, ..., x_n, ... | P_1(x), P_2(x), ..., P_s(x)\}$ мають властивості $P_1(x)$ і $P_2(x)$ і … і $P_s(x)$ та є неаномальними даними, елементи підмножини $Z = \{z_1, z_2, ..., z_j, ... | P_{a1}(z)$ або $P_{a2}(z)$ або...або $P_{j1}(z)\}$ у свою чергу не мають цих властивостей, що означає, що вони є аномальними у заданій множині даних А, а також, мають свої задані властивості $P_{a1}(z)$ або $P_{a2}(z)$ або … або $P_{j1}(z)$, відповідно $X \subseteq A$ та $Z \subseteq A$ та $X \cap Z = \varnothing$.

У свою чергу, якщо множина даних користувача міститиме елементи підмножини *Z*, то користувач вважатиметься шахраєм, інакше ж — органічним.

Для визначення властивостей підмножин неаномальних даних *X* та аномальних даних *Z* в області інсталювання мобільних додатків у роботі було проведено експертне опитування, в якому прийняли участь 26 експертів з Швейцарії, України, Ізраїлю, США [1], що мали досвід з виявлення аномалій в даних. Зібравши отримані характеристики даних, представимо деякі з них з використанням теорії множин. Так наприклад, деякі з властивостей підмножини аномальних даних *Z* вказано у формулах (2)—(6):

$$P_{a1}(z) = z \notin P, z.confirmedPurch = true, \quad (2)$$

де P — множина ідентифікаторів покупок, здійснених у мобільному додатку, *confirmed* — прапорець, що означатиме, підтверджена покупка зі сторони магазину чи ні.

$$P_{a2}(z) = z \notin R, \quad (3)$$

де R — множина відомих аномальних значень, яких може набувати елемент z. Так наприклад, у межах задачі, яка розглядається в роботі, це може бути завчасно відома множина шахрайських IP-адрес, множина ідентифікаторів відомих шахраїв (користувачів, які навмисно створювали аномальні дані), множина ідентифікаторів пристроїв, з яких спостерігалися аномалії. Ця властивість поділяється на такі:

$$P_{a2\_1}(z) = z \in IP\_FRAUD, z \notin R_{ip}, \quad (4)$$

де IP_FRAUD — множина IP-адрес користувачів, які заходили у мобільний додаток, що означа-

тиме, що цей елемент є IP-адресою, R_{ip} — множина завчасно відомих IP-адрес користувачів, які навмисно створювали аномальні дані;

$$P_{a2\_2}(z) = z \in ID\_FRAUD, z \notin R_{id}\,, \tag{5}$$

де ID_FRAUD — множина унікальних ідентифікаторів зареєстрованих у мобільному додатку користувачів, R_{id} — множина завчасно відомих ідентифікаторів користувачів, які навмисно створювали аномальні дані;

$$P_{a2\_3}(z) = z \in D\_FRAUD, z \notin R_{d}\,, \tag{6}$$

де D_FRAUD — множина унікальних ідентифікаторів пристроїв користувачів, з яких виконувався вхід у мобільний додаток; R_d — множина завчасно відомих ідентифікаторів пристроїв користувачів, які навмисно створювали аномальні дані.

Таке розділення даних на підмножини в подальшому дозволить розробити метод аналізу різнорідних даних та процедуру виявлення аномалій (шахраїв) при інсталюванні мобільних додатків з його використанням.

Виявлення повної інформації про користувача

Аналіз даних, які використовуються у системах виявлення шахрайства при інсталюванні мобільних додатків, показав, що даних дуже багато, але найскладнішим у їх використанні є те, що вони різнорідні [1], [6]. Під різнорідністю розуміється, що всі дані, а саме дані з різних класів (якісні, кількісні), неможливо порівняти між собою. Також слід зазначити, що серед наявних даних систем, що розглядаються, є не лише якісні і кількісні дані, але й масиви якісних та кількісних даних. Існуючі дані систем виявлення шахрайства показано на рис. 3, з якого видно, що всі дані є різнорідними. Багато існуючих систем-аналогів, методів та моделей відкидають багато даних, таких як, наприклад, масиви якісних даних чи дані, всі значення яких завчасно невідомі тощо через те, що з такими даними важко працювати і не завжди зрозуміло як їх аналізувати. Розглянемо приклад, чому це так. Так наприклад, такий якісний параметр як IP-адреса присутня у кожного користувача мобільного додатку. Деякі системи перевіряють наявність поточної IP-адреси у відомих базах даних з IP-адресами шахраїв та не подають її у якості ознаки (feature) в систему інтелектуального аналізу даних через те, що моделі інтелектуального аналізу даних в основному працюють лише з числовими значеннями; для перетворення якісних даних у кількісні зазвичай використовується метод one-hot encoding. Для використання цього методунеобхідна завчасно відома множина можливих значень ознаки (feature), у випадку з IP-адресою — завчасно невідома вся множина IP-адрес майбутніх користувачів мобільного додатку. При цьому, можливо згенерувати всі можливі значення IP-адрес, але це приведе до зберігання величезної кількості надлишкової інформації, погіршення ефективності та швидкодії роботи системи, тому зазвичай такі дані відкидаються і не подаються в модель прийняття рішення. Але ж від кількості вхідних даних, залежить кількість різних шаблонів шахраїв, які можна буде виділити, тобто чим більше даних, тим вища точність системи. Тому для правильного та ефективного використання повної інформації про користувача перейдемо до задачі аналізу даних з використанням методу аналізу різнорідних даних.

Аналіз різнорідних даних

З рис. 3 випливає, що всі вилучені дані є різнорідними, як і було зазначено вище. Отже, зрозуміло, що необхідно здійснити нормалізацію даних для приведення їх до однорідних. Для цього зазвичай застосовують шкали по даним у процесі їх аналізу. Проте на наш погляд, доцільно робити шкалювання по даним не в процесі їх аналізу, а по кінцевій меті поставленої задачі. Так наприклад, якщо ми матимемо бінарну шкалу, яка визначатиме, чи робив користувач покупку з непідтвердженим ідентифікатором, то по ній ми зможемо однозначно визначити, чи цей користувач є шахраєм. Аналогічно, можна ввести бінарну шкалу, яка однозначно визначатиме, чи користувач є шахраєм по його IP-адресі, а саме — якщо вже були раніше помічені шахраї з цієї IP-адреси, то поточний користувач також вважатиметься шахрайським. Також, якщо користувач використовує типи події, які йому ще недоступні, то він також є шахраєм. Розглянемо детальніше дані користувача та те, як їх можна об’єднати та прошкалювати, щоб створити висновок про те, чи ця інформа-

ція про шахрая чи ні. Для створення таких шкал проведено експертне опитування, у якому прийняли участь 26 експертів, що мали досвід у цій предметній області.

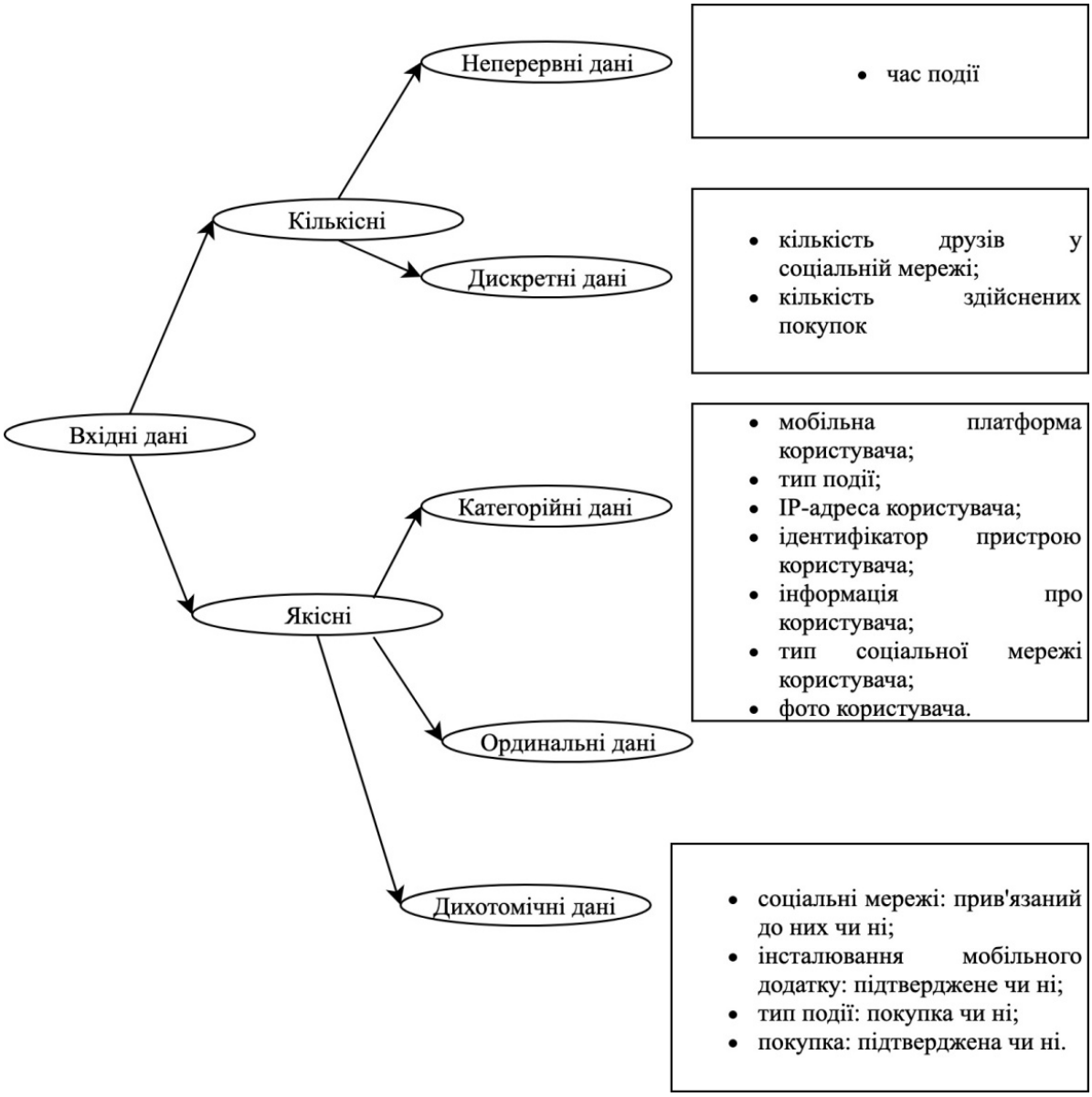

Рис. 3. Класифікація різнорідних вхідних даних

Експертами виділено шкали, які доцільно розділити на 2 групи [1]:

– група зі шкалами, по яким можна однозначно визначити чи користувач є шахраєм, чи органічним;

– група зі шкалами, по яким не можна однозначно визначити, до якого класу належить користувач, проте вони є важливими для прийняття рішення.

Вищерозглянуті шкали по ідентифікатору покупки та IP-адресі відносяться до першої групи, оскільки вони дозволяють однозначно визначити, до якого класу відноситься користувач. Зазначимо, що шкали першої групи дозволяють однозначно визначити клас користувача та сформувати базу даних та базу знань з відомими визначеними користувачами. До другої групи відносяться наприклад шкали по кількості друзів користувача у соціальній мережі чи середній час інсталювання. Експерти не можуть точно сказати ліміти для цих значень, проте з використанням попередньо сформованих бази даних та бази знань та коефіцієнтів подібності [5], [17] можна нормалізувати дані по шкалам другої групи.

Таким чином, для аналізу різнорідних даних з використанням запропонованих в роботі шкал та коефіцієнтів, отримано математичну модель процесу аналізу різнорідних даних, яка містить коефіцієнти визначеної метрики.

$$\vec{D}\begin{pmatrix} U_1(d_1,d_2,...,d_{p1}) \\ U_2(d_1,d_2,...,d_{p2}) \\ ..... \\ U_n(d_1,d_2,...,d_{pn}) \end{pmatrix} \rightarrow \begin{cases} \vec{G}_1\begin{pmatrix} U_1(g_{11},...,g_{1r}) \\ U_2(g_{11},...,g_{1r}) \\ ..... \\ U_n(g_{11},...,g_{1r}) \end{pmatrix} \rightarrow \\ \vec{G}_2\begin{pmatrix} U_1(g_{21},...,g_{2l}) \\ U_2(g_{21},...,g_{2l}) \\ ..... \\ U_n(g_{21},...,g_{2l}) \end{pmatrix} \rightarrow \end{cases} \begin{cases} \vec{X}\begin{pmatrix} U_1(x_1,...,x_n) \\ U_2(x_1,...,x_n) \\ ..... \\ U_n(x_1,...,x_n) \end{pmatrix} \rightarrow A_1(\vec{X}) \rightarrow \vec{X}_1\begin{pmatrix} U_n(k_{01}) \\ U_n(k_{02}) \\ ..... \\ U_n(k_{0n}) \end{pmatrix} \\ \vec{Y}\begin{pmatrix} U_1(y_1,...,y_k) \\ U_2(y_1,...,y_k) \\ ..... \\ U_n(y_1,...,y_k) \end{pmatrix} \rightarrow A_2(\vec{Y}) \rightarrow \vec{Y}_1\begin{pmatrix} U_n(k_{11}) \\ U_n(k_{12}) \\ ..... \\ U_n(k_{1n}) \end{pmatrix} \rightarrow \\ ..... \\ \vec{Z}\begin{pmatrix} U_1(z_1,...,z_m) \\ U_2(z_1,...,z_m) \\ ..... \\ U_n(z_1,...,z_m) \end{pmatrix} \rightarrow A_3(\vec{Z}) \rightarrow \vec{Z}_1\begin{pmatrix} U_n(k_{21}) \\ U_n(k_{22}) \\ ..... \\ U_n(k_{2n}) \end{pmatrix} \end{cases} \quad (7)$$

$$\rightarrow A_4(\vec{X}_1,\vec{Y}_1,\vec{Z}_1) \rightarrow \vec{V}\begin{pmatrix} U_1(k_{01},k_{11},...,k_{21}) \\ U_1(k_{02},k_{12},...,k_{22}) \\ ..... \\ U_n(k_{0n},k_{1n},...,k_{2n}) \end{pmatrix},$$

де $\vec{D}\begin{pmatrix} U_1(d_1,d_2,...,d_{p1}) \\ U_2(d_1,d_2,...,d_{p2}) \\ ..... \\ U_n(d_1,d_2,...,d_{pn}) \end{pmatrix}$ — дані по кожному з користувачів з бази даних мобільного додатку, а саме — вектор, що містить вектори з усіма визначеними ознаками по кожному з користувачів $(U_1,U_2,...,U_n)$;

$\vec{G}_1$ та $\vec{G}_2$ — вектори різнорідних вхідних даних першої та другої груп, поділені по принципу, визначеному на основі експертного опитування; $\vec{X}$,$\vec{Y}$,...,$\vec{Z}$ — вектори однорідних даних, поділених по типам; $A_1(\vec{X})$, $A_2(\vec{Y})$,…, $A_3(\vec{Z})$ — функції переводу векторів однорідних даних за певною ознакою у критерій, значення якого від 0 до 1 за відповідною шкалою. В основі даних функцій лежать відповідно визначені коефіцієнти. На виході даних функцій будуть вектори $\vec{X}_1,\vec{Y}_1,\vec{Z}_1$, які міститимуть користувачів зі значенням критерію по відповідній ознаці з використанням відповідної шкали; $A_4\left(\vec{X}_1,\vec{Y}_1,\vec{Z}_1\right)$ — функція об'єднання усіх значень по шкалам по користувачах у вектор однорідних даних $\vec{V}$.

Слід зазначити, що вектор однорідних даних $\vec{V}$ можна подати у класифікаційну модель для визначення класу кожного з користувачів.

На основі запропонованої математичної моделі процесу аналізу різнорідних даних (7) розроблено алгоритм аналізу різнорідних даних. Розглянемо його:

1. Розбиття вхідних даних $\vec{D}\begin{pmatrix} U_1(d_1,d_2,...,d_{p1}) \\ U_2(d_1,d_2,...,d_{p2}) \\ ..... \\ U_n(d_1,d_2,...,d_{pn}) \end{pmatrix}$ на дві групи $\vec{G}_1$ та $\vec{G}_2$ (7). Зазначимо, що дані групи $\vec{G}_1$ дозволяють визначити однозначних органічних та шахрайських користувачів, а група $\vec{G}_2$ для шкалювання потребує використання коефіцієнтів подібності між користувачами.

2. Розбиття даних першої групи $\vec{G}_1$ на вектори однорідних даних.

3. Визначення коефіцієнтів першої групи даних $\vec{G}_1$ за допомогою здійснення процедури шкалювання даних у коефіцієнти. Значення коефіцієнтів будуть бінарними, а саме — 0 або 1. Для цього необхідно здійснити:

– шкалювання якісних даних, що зазвичай залежить від їх властивостей, які зазвичай задані належністю елемента до завчасно відомої вибірки.

– шкалювання кількісних даних, що зазвичай залежить від їх властивостей, заданих граничними значеннями.

– шкалювання множини даних, що зазвичай залежить від їх властивостей, заданих належністю множини до певного відомого закону розподілу.

4. Визначення однозначно відомих шахраїв та органічних користувачів по кожному з векторів однорідних даних та створення загальної множини органічних користувачів та множини користувачів, які навмисно створювали аномальні дані (шахраїв).

5. Розбиття даних другої групи $\vec{G}_2$ на вектори однорідних даних.

6. Визначення коефіцієнтів подібності невизначених користувачів, для цього необхідно здійснити:

– визначення коефіцієнтів подібності невизначених користувачів на основі векторів однорідних даних користувачів та користувачів з множини.

– визначення коефіцієнтів подібності невизначених користувачів на основі векторів однорідних даних користувачів та користувачів з множини.

Вибір коефіцієнтів подібності в залежності від типу ознаки розглянуто в роботі [5].

7. Формування нових характеристик органічних користувачів та шахраїв на основі пункту 6 та доповнення множин з однозначно визначеними користувачами-шахраями та органічними користувачами, відповідно.

8. Отримання вектора однорідних даних, який можна подати у модель класифікації.

Приклад перетворених однорідних даних, які подаватимуться у модель класифікації зображено на рис. 5 (вхідні дані взято з бази даних дій користувачів мобільного додатку, наданої представниками IT-компанії).

```
appActions.head()
```

Out[2]:

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...
0	0.0	0.359807	0.072781	0.536347	0.378155	0.338321	0.462388	0.239599	0.098698	0.363787	...
1	0.0	0.191857	0.266151	0.166480	0.448154	0.060018	0.082361	0.078803	0.085102	0.255425	...
2	1.0	0.358354	0.340163	0.773209	0.379780	0.503198	0.800499	0.791461	0.247676	0.514654	...
3	1.0	0.966272	0.185226	0.792993	0.863291	0.010309	0.247203	0.237609	0.377436	0.387024	...
4	2.0	0.158233	0.877737	0.548718	0.403034	0.407193	0.095921	0.592941	0.270533	0.817739	...

Рис. 5. Екранна форма роботи програми з однорідними данними, отриманими за використання методу аналізу різнорідних даних по кожному користувачу

Запропоновані метод та алгоритм використані під час розробки методу аналізу різнорідних даних та інтелектуальної системи виявлення аномалій при інсталюванні мобільних додатків.

Так, після такого перетворення даних з використанням методу аналізу різнорідних даних можна побачити, що всі значення отриманих ознак мають числові значення, нормалізовані у межах [0; 1]. А такі однорідні дані можна подати у класифікаційну модель, таку як XGBoost або випадковий ліс (random forest), щоб отримати остаточну класифікацію. Для вирішення задачі класифікації авторами обрано повністю зв'язану глибоку нейронну мережу (fully-connected DNN) з 3 прихованими шарами для отримання високої точності (99,14 %). Отже, експериментальні дані показують, що висока продуктивність та точність системи виявлення шахрайства отримана в основному за рахунок перетворення різнорідних даних в однорідні (запропонований метод аналізу різнорідних даних) і завдяки посиланню на сформовані узагальнені шаблони органічних і шахрайських користувачів. А саме, на основі сформованих бази знань та бази даних формується узагальнений шаблон (портрет) шахрая, який дозволяє для нових великих масивів даних визначати шахраїв з точністю 99,14 % та вирішує останню поставлену задачу. Також слід зазначити, що наявність сформованих бази знань, бази даних та узагальненого шаблону шахрая дозволяє прискорити процес виявлення шахраїв у нових наборах даних та виявляти навіть неявних шахраїв.

Експериментальні дослідження

Для експериментального дослідження запропонованого методу у межах виявлення шахрайства як аномалій в різнорідних даних розроблено програмне забезпечення [1], [18], [19], в основі якого лежать запропоновані у роботі математична модель, алгоритм та метод аналізу різнорідних даних.

Схему експериментального дослідження виявлення аномалій в різнорідних даних при інсталюванні мобільних додатків, в основі якої є представлений метод аналізу різнорідних даних подано на рис. 6.

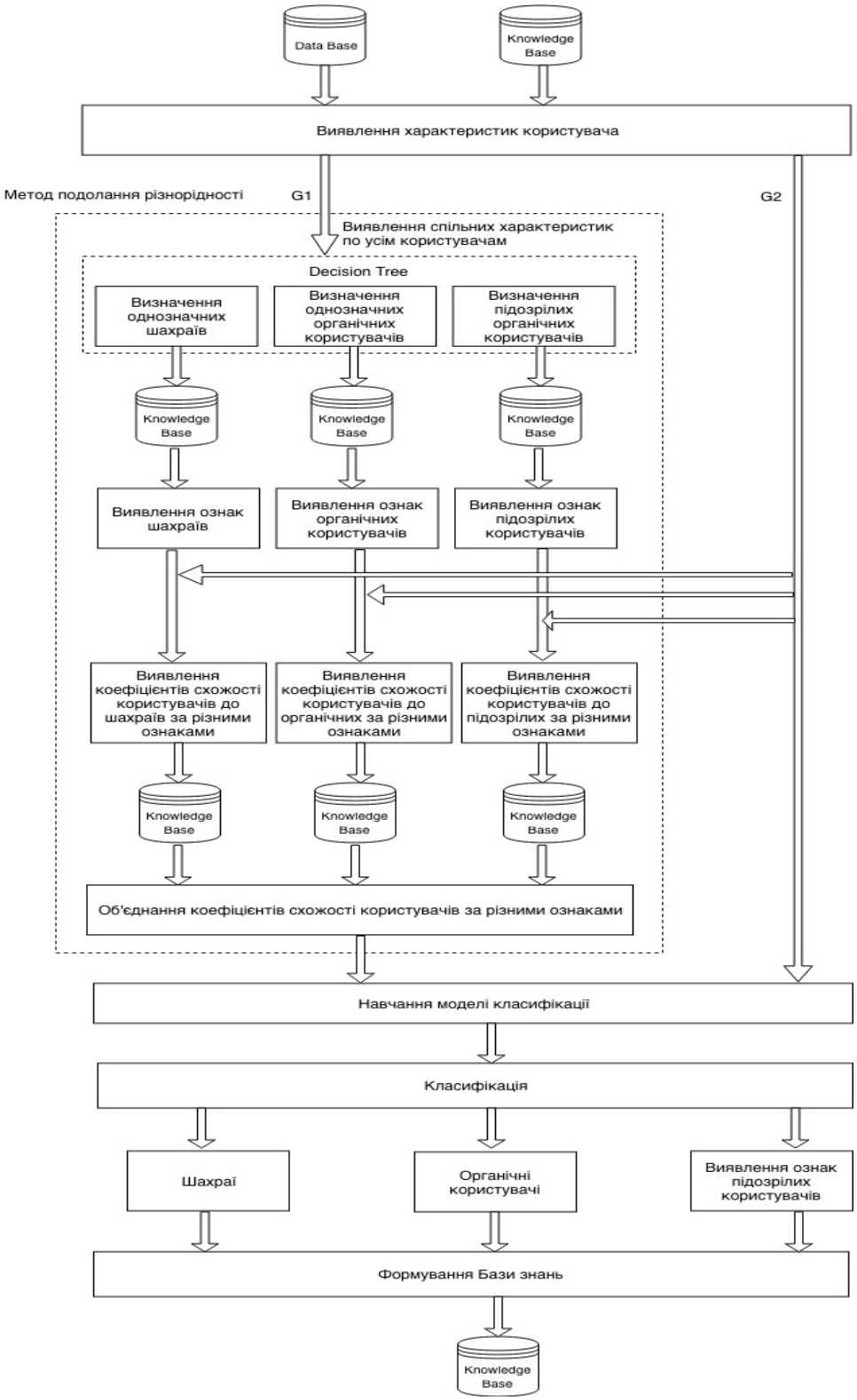

Рис. 6. Схема експериментального дослідження виявлення аномалій в різнорідних даних при інсталюванні мобільних додатків з використанням методу аналізу різнорідних даних

Дослідження проводилося в такому порядку:

1. Отримання даних з бази даних та бази знань користувачів.

2. Виявлення характеристик користувача та розбиття даних на дві групи $\vec{G}_1$ та $\vec{G}_2$.

3. Метод аналізу різнорідних вхідних даних, виявлення спільних характеристик по усім користувачам та формування вектору однорідних даних по кожному з користувачів.

5. Навчання моделі класифікації та класифікація користувачів на отриманих однорідних даних та з використанням повністю зв'язаної глибокої нейронної мережі з трьома прихованими шарами. Перевірка результатів на поміченому наборі даних, про що детальніше описано в [6].

7. Визначення шахраїв та органічних користувачів.

8. Формування бази знань з використанням нечіткої логіки (детальніше описано в [5]).

Результати експерименту подано на рис. 7. Загальна точність моделі класифікації, яка працює з даними, отриманими з методу аналізу різнорідних даних, 99,14 %, точність виявлення шахраїв — 82,76 %. Проте зазначимо, що зі збільшенням правил у базі знань, що буде відбуватися з кожним запуском на нових даних, збільшуватиметься точність.

```
Percent of fraudulent transactions:  0.001727485630620034
/Users/tetianapolhul/PycharmProjects/CreditCardFraudDetecti
  return f(*args, **kwds)
2018-07-24 23:33:11.755811: I tensorflow/core/platform/cpu_
Epoch: 0 Current loss: 1.4053 Elapsed time: 1.58 seconds
Current accuracy: 0.15%
Epoch: 10 Current loss: 1.4053 Elapsed time: 1.32 seconds
Current accuracy: 0.15%
Epoch: 20 Current loss: 1.3875 Elapsed time: 1.20 seconds
Current accuracy: 0.15%
Epoch: 30 Current loss: 1.3002 Elapsed time: 1.25 seconds
Current accuracy: 66.30%
Epoch: 40 Current loss: 1.1396 Elapsed time: 1.21 seconds
Current accuracy: 93.02%
Epoch: 50 Current loss: 1.0138 Elapsed time: 1.21 seconds
Current accuracy: 97.49%
Epoch: 60 Current loss: 0.9332 Elapsed time: 1.29 seconds
Current accuracy: 99.00%
Epoch: 70 Current loss: 0.8944 Elapsed time: 1.14 seconds
Current accuracy: 99.46%
Epoch: 80 Current loss: 0.8729 Elapsed time: 1.33 seconds
Current accuracy: 99.65%
Epoch: 90 Current loss: 0.8608 Elapsed time: 1.16 seconds
Current accuracy: 99.62%
Final accuracy: 99.14%
Final fraud specific accuracy: 82.76%

Process finished with exit code 0
```

Рис. 7. Результати комп'ютерного моделювання, здійсненого на основі даних, отриманих з методу аналізу різнорідних вхідних даних

Висновки

Таким чином, у роботі формалізовано процес виявлення шахрайства як аномалії в даних при інсталюванні мобільних додатків, виявлено повну інформацію про користувача на основі проведеного аналізу даних, які використовуються у системах виявлення шахрайства при інсталюванні мобільних додатків. Запропоновано підхід до аналізу всіх наявних різнорідних даних різних шаблонів, розмірності, метрик з використанням розробленої математичної моделі, алгоритму та методу аналізу різнорідних даних. Це дозволило використати інформацію про усі наявні вхідні дані по користувачам, значно підвищити ефективність процедури виявлення нових шахрайських користувачів та отримати високу точність вирішення поставленої задачі.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] T. Polhul, and A. Yarovyi "Development of a method for fraud detection in heterogeneous data during installation of mobile applications," *Eastern-European Journal of Enterprise Technologies*, № 1/2 (97), 2019. https://doi.org/doi: 10.15587/1729-4061.2019.155060

[2] D. Hawkins, "Identification of Outliers," *Chapman and Hall*, 1980.

[3] А. А. Яровий, О. Н. Романюк, І. Р. Арсенюк, та Т. Д. Польгуль, «Виявлення шахрайства при інсталюванні програмних додатків з використанням інтелектуального аналізу даних,» *Наукові праці Донецького національного технічного університету. Серія: «Інформатика, кібернетика та обчислювальна техніка»*, № 2 (25), с. 126-131, 2017. [Електронний ресурс]. Режим доступу: http://science.donntu.edu.ua/wp-content/uploads/2018/03/ikvt_2017_2_site-1.pdf .

[4] Т. Д. Польгуль, та А. А. Яровий, «Визначення шахрайських операцій при встановленні мобільних додатків з використанням інтелектуального аналізу даних,» *Сучасні тенденції розвитку системного програмування. Тези доповідей.* Київ, 2016. с. 55-56. [Електронний ресурс]. Режим доступу: http://ccs.nau.edu.ua/wp-content/uploads/2017/12/%D0%A1%D0%A2%D0%A0%D0%A1%D0%9F_2016_07.pdf .

[5] Т. Д. Польгуль, та А. А. Яровий, «Метод подолання різнорідності даних для виявлення шахрайства при інсталюванні мобільних додатків,» *Вісник СНУ ім. В. Даля*, № 7 (248) с. 60-69, 2018.

[6] T. Polhul, "Development of an intelligent system for detecting mobile app install fraud," *Proceedings of the IRES 156th International Conference*, Bangkok, Thailand, 21st-22nd March 2019. pp. 25-29.

[7] Kochava Uncovers Global Ad Fraud Scam. [Електронний ресурс]. Режим доступу: https://www.kochava.com/ .

[8] Andrii Yarovyi, Raisa Ilchenko, Ihor Arseniuk, Yevhene Shemet, Andrzej Kotyra, and Saule Smailova, "An intelligent system of neural networking recognition of multicolor spot images of laser beam profile". *Proc. SPIE 10808, Photonics Applications in Astronomy, Communications*, Industry, and High-Energy Physics Experiments 2018, 108081B (1 October 2018). https://doi.org/10.1117/12.2501691 .

[9] V. Kozhemyako, L. Timchenko, and A. Yarovyy, "Methodological Principles of Pyramidal and Parallel-Hierarchical Image Processing on the Base of Neural-Like Network Systems," *Advances in Electrical and Computer Engineering*, vol. 8, no. 2, pp. 54-60, 2008, https://doi.org/10.4316/AECE.2008.02010 .

[10] M. Granik, V. Mesyura and A. Yarovyi, "Determining Fake Statements Made by Public Figures by Means of Artificial Intelligence," *2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT)*, Lviv, 2018, pp. 424-427. https://doi.org/ 10.1109/STC-CSIT.2018.8526631 .

[11] R. Agrawal, R. Srikant, "Mining sequential patterns," *Proceedings of the Eleventh International Conference on Data Engineering. 1995.* doi: https://doi.org/10.1109/icde.1995.380415

[12] V. Chandola, A. Banerjee, V. Kumar, "Anomaly detection," *ACM Computing Surveys*, vol. 41, iss. 3, pp. 1-582009. https://doi.org/https://doi.org/10.1145/1541880.1541882 .

[13] S. Guido, A. Müller, *Introduction to Machine Learning with Python: A Guide for Data Scientists.* O'Reilly Media, 2016. 400 p.

[14] D.-Y. Yeung, C. Chow, "Parzen-window network intrusion detectors," "Object recognition supported by user interaction for service robots." 2002. https://doi.org/10.1109/icpr.2002.1047476 .

[15] E. Keogh, J. Lin, A. Fu, "HOT SAX: Efficiently Finding the Most Unusual Time Series Subsequence," *Fifth IEEE International Conference on Data Mining (ICDM'05).* 2005. https://doi.org/10.1109/icdm.2005.79 .

[16] E. Keogh, J. Lin, S.-H. Lee, H. V. Herle "Finding the most unusual time series subsequence: algorithms and applications," *Knowledge and Information Systems*, vol. 11, iss. 1, pp. 1-27 , 2006. https://doi.org/10.1007/s10115-006-0034-6 .

[17] А. Г. Кюльян, Т. Д. Польгуль, та М. Б. Хазін, «Математична модель рекомендаційного сервісу на основі методу колаборативної фільтрації,» *Комп'ютерні технології та Інтернет в інформаційному суспільстві*, с. 226-227, 2012. [Електронний ресурс] Режим доступу: http://ir.lib.vntu.edu.ua/bitstream/handle/123456789/7911/226-227.pdf?sequence=1&isAllowed=y

[18] А. А. Яровий, та Т. Д. Польгуль, «Комп'ютерна програма «Програмний модуль збору даних інформаційної технології» виявлення шахрайства при інсталюванні програмних додатків.» *Свідоцтво про реєстрацію авторського права на твір № 76348.* К.: Міністерство економічного розвитку і торгівлі України, 2018.

[19] А. А. Яровий, та Т. Д. Польгуль, «Комп'ютерна програма «Програмний модуль визначення схожості користувачів інформаційної технології виявлення шахрайства при інсталюванні програмних додатків,» *Свідоцтво про реєстрацію авторського права на твір № 76347.* К.: Міністерство економічного розвитку і торгівлі України, 2018.

Рекомендована кафедрою комп'ютерних наук ВНТУ

Стаття надійшла до редакції 15.04.2019

Польгуль Тетяна Дмитрівна — аспірантка кафедри комп'ютерних наук, e-mail: tanapolg93@gmail.com ;

Яровий Андрій Анатолійович — д-р техн. наук, професор, завідувач кафедри комп'ютерних наук, e-mail: a.yarovyy@vntu.edu.ua .

Вінницький національний технічний університет, Вінниця

T. D. Polhul[1]
A. A. Yarovyi[1]

Heterogeneous Data Analysis in Intelligent Fraud Detection Systems

[1]Vinnytsia National Technical University

Fraud is being considered as an anomaly in the data in the work. The work is devoted to the development of a method of heterogeneous data analysis in intelligent fraud detection systems. The detection of fraud as an anomaly in heterogeneous data during mobile applications installation using set theory, which allowed further data analysis in such systems, is formalized. The mathematical model of the process of heterogeneous data analysis, the algorithm of heterogeneous data analysis, the method of heterogeneous data analysis on the basis of the proposed scales and coefficients that allowed processing of various input data — data of various metrics, templates, dimensions, which in the analysis process makes it possible to form a generalized fingerprint of fraudster, is proposed. The developed method uses the databases and knowledge bases, through which a generalized fingerprint of the fraudster is formed, the presence of which allows accelerating the detection of fraudsters in new data sets and detecting even implicit fraudsters. The proposed method is designed to use it in intelligent systems for fraud detection based on anomalies in data that, unlike existing ones, will allow analyzing heterogeneous data

on the basis of which fraudulent decisions are made, to reduce the dimensionality of data and to classify users. Experimental researches of the proposed method of heterogeneous data analysis as a part of detection of fraud as anomalies in heterogeneous data and a classification model developed using fully connected deep neural networks with three hidden layers using the developed software and using a representative sample have been carried out. The scheme of experimental research of detection of anomalies in heterogeneous data during the mobile applications installation, based on which method of heterogeneous data analysis was presented has been proposed. The efficiency of using the proposed method in the fraud detection system is shown, the classification accuracy of which was 99,14 %, the accuracy of the fraud detection is 82,76 %. However, with the increase of rules in the developed knowledge base, which will grow with each launch on the new data, the accuracy of the system will increase.

Keywords: fraud detection, anomaly detection, classification model, method for analyzing heterogeneous data.

Polhul Tetiana D. — Post-Graduate student of the Chair of Computer Science, e-mail: tanapolg93@gmail.com ;
Yarovyi Andrii A. — Dr. Sc. (Eng.), Professor, Head of the Chair of Computer Science, e-mail: a.yarovyy@vntu.edu.ua

Т. Д. Польгуль[1]
А. А. Яровой[1]

Анализ разнородных данных в интеллектуальных системах обнаружения мошенничества

[1]Винницкий национальный технический университет

Авторами мошенничество рассматривается как аномалия в данных. Разработан метод анализа разнородных данных в интеллектуальных системах обнаружения мошенничества. Формализован процесс выявления мошенничества как аномалии в разнородных данных при инсталлировании мобильных приложений с использованием теории множеств, что позволило осуществить дальнейший анализ данных в таких системах. Предложена математическая модель процесса анализа разнородных данных, алгоритм анализа разнородных данных, метод анализа разнородных данных на основе предложенных шкал и коэффициентов, которые позволили обрабатывать разноформатные входные данные — различные метрики, шаблоны, размерности, что в процессе анализа дает возможность сформировать обобщенный шаблон мошенника. Разработанный метод использует базы данных и базы знаний, благодаря которым формируется обобщенный шаблон мошенника, наличие которого позволяет ускорить выявление мошенников в новых наборах данных и выявлять даже неявных мошенников. Предложенный метод разработан с целью его использования в интеллектуальных системах обнаружения мошенничества на основе аномалии в данных, которые, в отличие от существующих, позволят провести анализ разнородных данных, на основе которых принимаются решения о мошенничестве, уменьшить размерности данных и провести классификацию пользователей. Проведены экспериментальные исследования предложенного метода анализа разнородных данных в рамках выявления мошенничества как аномалии в разнородных данных и классификационной модели, разработанной с использованием полностью связанных глубоких нейронных сетей с тремя скрытыми слоями с использованием разработанного программного обеспечения, репрезентативной выборки. Предложена схема экспериментального исследования для выявления аномалий в разнородных данных при инсталлирования мобильных приложений, в основе которой лежит представленный метод анализа разнородных данных. Показана эффективность использования предложенного метода в системе определения мошенничества, точность классификации которой составила 99,1 4%, точность обнаружения мошенников — 82,76 %. Однако с увеличением правил в разработанной базе знаний, которое будет происходить с каждым запуском на новых данных, будет увеличиваться точность системы.

Ключевые слова: определение мошенничества, определение аномалий, модель классификации, метод анализа разнородных данных.

Польгуль Татьяна Дмитриевна — аспирант кафедры компьютерных наук, e-mail: tanapolg93@gmail.com ;
Яровой Андрей Анатолиевич — д-р техн. наук, профессор, заведующий кафедрой компьютерных наук, e-mail: a.yarovyy@vntu.edu.ua