УДК 004.94

МЕТОД ПОДОЛАННЯ РІЗНОРІДНОСТІ ДАНИХ ДЛЯ ВИЯВЛЕННЯ ШАХРАЙСТВА ПРИ ІНСТАЛЮВАННІ МОБІЛЬНИХ ДОДАТКІВ

Польгуль Т. Д., Яровий А. А.

THE INPUT DATA HETEROGENEITIES RESOLUTION METHOD DURING MOBILE APPLICATIONS INSTALLATION FRAUD DETECTION /

Polhul T., Yarovyi A.

У роботі запропоновано метод та алгоритми подолання різнорідності даних для виявлення шахрайства при інсталюванні мобільних додатків. Процедура виявлення шахрайства на основі розробленого методу дозволяє виявити шахраїв та визначити їх характеристики і шаблони. Здійснено експериментальні дослідження на основі обраної поміченої вибірки. В експериментах мітки класів не використовувались, проте вони необхідні для перевірки точності процедури прийняття рішень, розробленої на основі запропонованого методу, що склала 99,14 %.

Ключові слова: *виявлення шахрайства, подолання різнорідності, інтелектуальний аналіз даних, виявлення аномалій, інсталювання мобільних додатків.*

Постановка проблеми. Наявність шахрайства при інсталюванні мобільних додатків стала досить поширеною та вагомою (дорогою) проблемою. Так, наприклад, дослідження AppLift «Fighting Mobile Fraud in the Programmatic era» [1] показало, що загалом, 34% мобільного трафіку, за яким спостерігали, був шахрайським. У коштах це представляє більше ніж $4,5 млрд втрат.

Задачею шахрайства при інсталюванні мобільних додатків є внесення даних у додаток шахрайськими способами. Внесені дані шахраї намагаються зробити якомога більш наближеними до дій органічних користувачів. Оскільки компанії-розробники мобільних додатків витрачають кошти на маркетингові компанії, що у свою чергу повинні привести органічних користувачів, то метою такого шахрайства є виведення коштів з компаній-розробників без приведення органічних користувачів у відповідь. На цьому етапі слід виділити наступні відомі способи шахрайства при інсталюванні мобільних додатків:

- мобільне викрадення (mobile hijacking) [2 – 3] - відбувається, коли справжній мобільний додаток виконує деякі несанкціоновані дії. Наприклад, запускає приховані оголошення і формує кліки від імені користувача у фоновому режимі. У цілому, програма у цьому випадку буде працювати за сценарієм, що максимально імітуватиме поведінку людини;

- кліковий спам (click spamming) [2 – 5]. Цей спосіб відноситься до програм, які генерують підроблені запити кліків програмним способом;

- ферми дій (action farms) [2 – 3]. Шахраї винагороджують людей по всьому світу за інсталювання мобільних додатків у ручному режимі, тобто фактично відбувається найняття людей для того, щоб вони інсталювали мобільні додатки.

На даний момент існують аналоги виявлення шахрайства, так, наприклад, Kraken [4], Fraudlogix [5], Appsflyer [6], які ще називають системами для боротьби з шахрайством (anti fraud systems). Проте вирішення задачі виявлення шахрайства все одно можна вважати частковим, оскільки більшість існуючих систем-аналогів працюють наступним чином:

- визначають шахраїв неповністю, оскільки використовують готові бази даних IP-адрес користувачів, так наприклад Appsflyer [6] чи Kraken [4]. Через такий підхід деякі шахраї будуть невизначені;

- більшість систем працюють по жорстким алгоритмам або за жорсткими правилами. Відсоток виявлення шахраїв з використанням таких систем недостатньо високий, оскільки кількість шахраїв та їх характеристик може зростати з кожним інсталюванням, а жорсткі правила, у свою чергу, не будуть це враховувати;

- використовують не всі вхідні дані, оскільки на даний момент існує мало сучасних алгоритмів подолання різнорідності усіх типів даних. Невраховування деяких вхідних даних також призводить до невизначення деяких з шахраїв. Під різнорідністю даних розуміємо те, що дані можуть бути як кількісними і якісними, так і множинами кількісних та якісних даних. Якісні дані у свою чергу можуть бути з неви-

значеною множиною категорій-значень, а сучасні алгоритми подолання різнорідності працюють лише з визначеною кількістю категорій якісних даних.

Тобто ефективність таких систем недостатньо висока, тому існує потреба розробки автоматизованої системи виявлення шахрайства, яка використовує нові методи та алгоритми на базі інтелектуального аналізу даних, що дозволить знаходити не лише явних шахраїв, але й виявляти їх шаблони та характеристики і створювати бази даних та бази знань шахраїв для подальшого і ефективного їх використання.

Отже, враховуючи складність і різнорідність вхідних даних для систем, що розглядаються, дана робота присвячена розробці методу подолання різнорідності даних з метою підвищення ефективності виявлення шахрайства при інсталюванні мобільних додатків та нових шахрайських і органічних шаблонів, характеристик і залежностей.

Аналіз літератури. Розглянемо найбільш відомі та популярні системи виявлення шахрайства при інсталюванні та відслідковуванні дій мобільних додатків. Алгоритми пошуку шахраїв з більшості систем-аналогів не розкриваються розробниками, проте за отриманими результатами зрозумілі їх основні рішення. Отже, розглянемо відомі характеристики, переваги та недоліки деяких з них:

- система Fraudlogix [7] має лише чорний список IP-адрес, тобто аналізує користувачів лише по одному критерію. Недоліком даної системи є те, що вона не враховує появу нових шахраїв, ботів з новими параметрами та характеристиками та не враховує усі дані користувача;

- російська розробка Kraken [6] перевіряє належність джерела конверсії до будь-якої з відомих ферм ботів, через що має такий же недолік, як і попередня система. Також дана система не масштабується і робить перевірку на шахрайство лише вибірково по визначеним дням тижня, чим упускає велику кількість шахраїв;

- Adjust [9] використовує не всю множину вхідних даних, а перевіряє лише наявність IP-адреси в базі даних VPN, вивчає кількість однакових кліків з одного джерела та порівнює час між подіями. Використання більшої кількості вхідних даних дозволило б більш точно зробити оцінку користувачів та виявити шахраїв;

- Kochava [10] виявляє шахраїв лише за визначеними критеріями. Критеріїв підібрано достатньо багато, але недоліком такого підходу є те, що у шахраїв з'являються нові характеристики і нові дані, які кожного разу необхідно оновлювати. Також, дана система не використовує усі наявні дані користувача, що знижує її ефективність виявлення шахрайства;

- TMC Attribution Analytics [11] також використовує не всі вхідні дані користувача та має обмежену кількість критеріїв, за якими визначається шахрайство, через що має такі ж недоліки, як і попередня система;

- відома система виявлення шахрайства Fraud-Shield [12] має досить зручний інтерфейс та можливість налаштування кожного критерію. Проте саме власне налаштування вносить недоліки у дану систему, оскільки через деякі її налаштування можна або пропустити дуже багато шахраїв, або навпаки, через деякі налаштування система вважатиме всіх користувачів шахрайськими;

- ще одна існуюча система Forensiq [13] є достатньо відомою, кожну конверсію характеризує відповідним рівнем ризику (низький, середній, високий) та перерахуванням підозрілих ознак. Проте неможливо дізнатись чи відстежити, як система визначає рівень ризику та оцінку. Часто ці знання є важливими, особливо, у випадках, коли необхідно довести причину визначення користувачів, приведених певною маркетинговою кампанією, як шахрайських при судових позовах;

- Appsflyer [8] є провідною платформою мобільної атрибуції і маркетингової аналітики, яка не так давно випустила власну систему захисту від шахраїв Protect360 [14]. Система має величезну базу IP-адрес та пристроїв, які позначені міткою шахрайства. Як говорить сама компанія, визначена ними мітка шахрайства – це лише привід для додаткової експертизи, оскільки не завжди вона є коректною. Також, неможливо дізнатися причину позначення користувача шахраєм, важливість даної можливості вказана у попередній системі;

- FraudScore [15] є системою, яка на відміну від більшості має оновлення своїх алгоритмів та має самонавчальну систему на базі нейронної мережі. Проте усі свої оцінки вона робить на основі лише деяких даних (таких як дані про пристрій користувача, геодані, дані, пов'язані з операційною системою тощо). Але аналіз не усіх даних також спричиняє невиявлення деяких шахраїв;

- AppMetrica [16] у свою чергу просто провела інтеграцію з попередньо розглянутою системою FraudScore, а отже має ті самі характеристики та недоліки, що і попередня система.

Слід зазначити, що більшість розглянутих систем, використовуються не лише для виявлення шахрайства при інсталюванні мобільних додатків, а є узагальненими для всіх мобільних транзакцій. Це спричиняє невикористання специфічних даних, які притаманні саме мобільним додаткам, що знижує ефективність.

Використання усіх даних користувачів підвищило б ефективність виявлення шахрайства, завдяки знаходженню усіх можливих нових шаблонів, характеристик та залежностей шахраїв з використанням інтелектуального аналізу даних. Проте для використання усіх даних необхідно вирішити задачу подолання їх різнорідності.

Мета статті. Розробка методу подолання різнорідності даних при виявленні шахрайства під час інсталюванні мобільних додатків, який на відміну від існуючих використовує усю множину вхідних даних, що дозволяє знаходити нові шахрайські шаб-

лони, їх характеристики та залежності для виявлення як явних, так і неявних шахраїв та підвищення ефективності процесу виявлення шахрайства при інсталюванні мобільних додатків вцілому.

Аналіз даних при інсталюванні мобільних додатків. З метою автоматизації процесу виявлення шахрайства та розробки методу подолання різнорідності, що лежатиме в основі процесу, здійснимо аналіз всіх даних про користувачів, на основі яких приймається рішення про наявність шахраїв. Для цього спочатку проаналізуємо ці дані. Зазначимо, що множина усіх подій користувача, яку необхідно проаналізувати для виявлення шахрайства, поділятиметься на:

- множину подій до реєстрації;
- множину внутрішніх подій додатку;
- множину подій після використання додатку.

У свою чергу, типи подій та дані по кожному з них по кожній з множин представлено на рисунках 1 – 3 відповідно.

Множина подій користувача до реєстрації містить такі типи подій як:

- подія інсталювання додатку. Даний тип події супроводжується такими даними як ідентифікатор користувача, IP-адреса користувача, ідентифікатор пристрою користувача, операційна система пристрою користувача та час події;
- подія отримання підтвердження про інсталювання додатку, яка супроводжується такими ж даними як і подія інсталювання додатку, а також має бінарну відмітку про те, чи інсталювання було підтверджене, чи ні;
- подія відкриття користувачем додатку супроводжується усіма спільними з попередніми даними.

Множина внутрішніх подій додатку, які відрізняються від попередніх наступними полями, така:

- подія про реєстрацію користувача – крім спільних зі всіма подіями даних, містить інформацію про користувача, фото користувача, дату народження користувача;

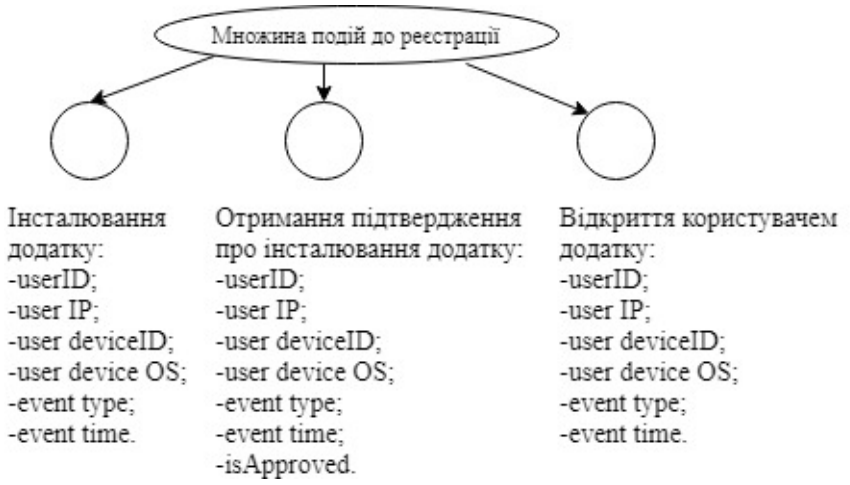

Рис. 1. Множина типів подій користувача та їх дані до реєстрації мобільного додатку

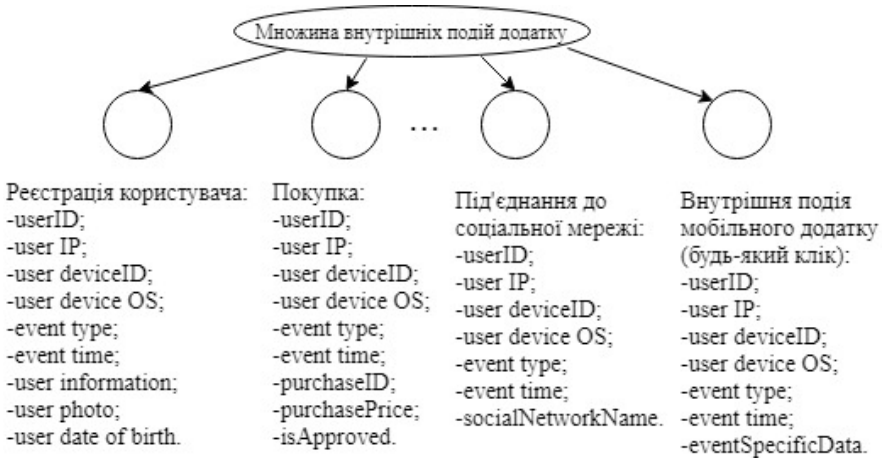

Рис. 2. Множина типів внутрішніх подій додатку та їх дані

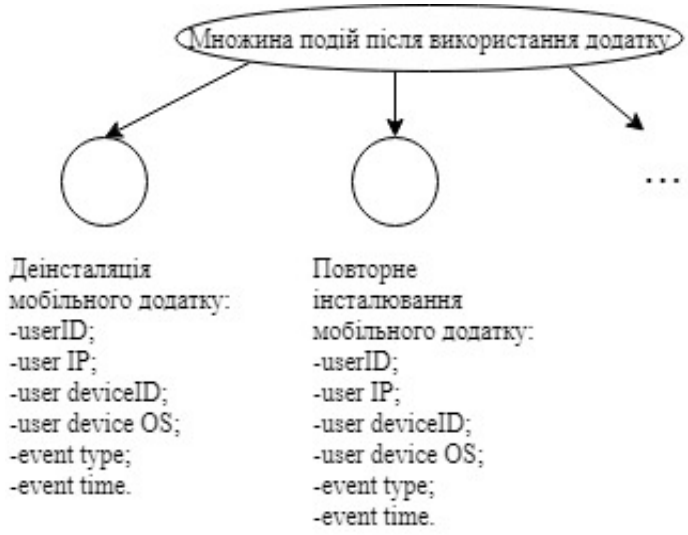

Рис. 3. Множина типів подій користувача після використання додатку та їх дані

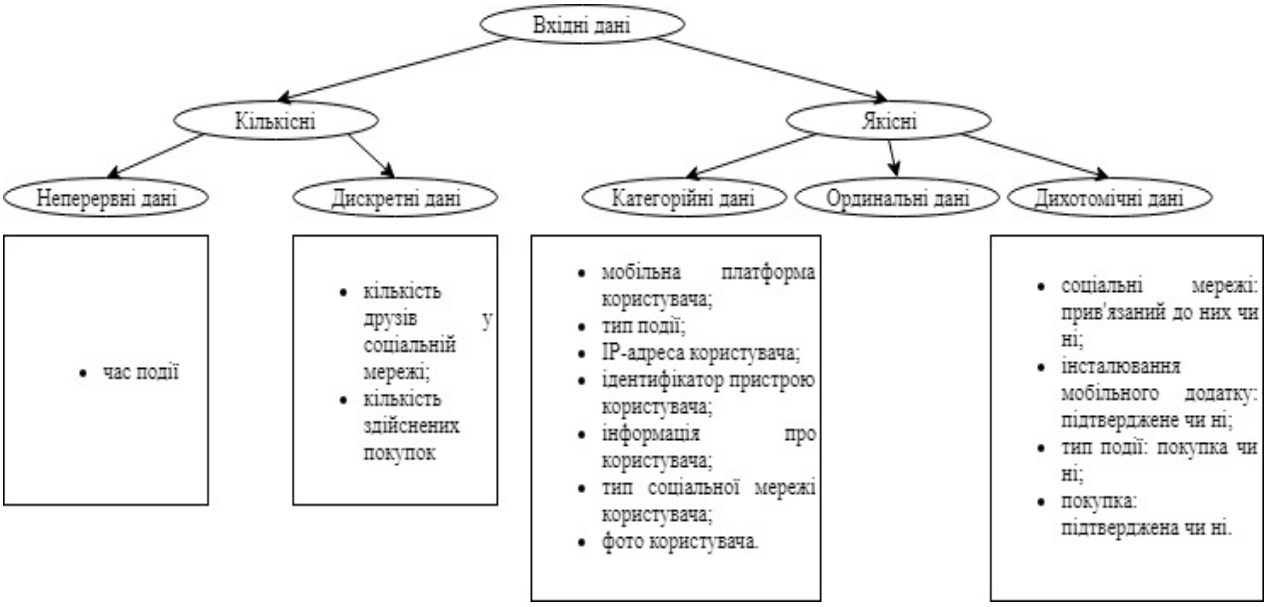

Рис. 4. Класифікація вхідних даних мобільного додатку

- подія про покупку, крім спільних даних, також містить ідентифікатор покупки, ціну покупки та бінарну ознаку, що показує, чи покупка була підтверджена магазином за вказаним ідентифікатором;

- подія, що сповіщує про під'єднання користувачем до соціальної мережі, містить додаткову інформацію про назву соціальної мережі;

- кожна з внутрішніх подій додатку містить специфічну для поточної події інформацію.

Множина подій після використання додатку наступна:

- подія про деінсталювання мобільного додатку;

- подія про повторне інсталювання мобільного додатку.

Події даної множини не містять особливих даних.

Відповідно, дані по кожному з типів подій, які показано на рисунках 1 – 3 можна характеризувати за групами, представленими на рисунку 4.

На рисунку 4 показано різнорідні дані, що є в наявності для виявлення шахрайства при інсталюванні мобільних додатків. Розглянемо їх більш детально:

- до кількісних даних відноситься час події, що присутній у всіх типах подій, кількість друзів у соціальній мережі та кількість здійснених покупок. Зазначимо, що розглянуті системи-аналоги не використовують дані з соціальних мереж та дані про кількість здійснених покупок при виявленні шахрайства;

- існуючі якісні дані можна поділити на категорійні та дихотомічні. До категорійних даних відноситься назва мобільної платформи користувача, тип події, IP-адреса користувача, ідентифікатор пристрою користувача, інформація про користувача,

тип соціальної мережі та фото користувача. Слід зазначити, що існуючі методи переведення якісних даних (ознак) у кількісні, такі наприклад як one-hot encoding [17], працюють лише тоді, коли існує відома кількість значень категорій даної ознаки. Проте якщо кількість значень мобільної платформи та типу події може бути однозначно визначена, то кількість значень усіх інших ознак неможливо однозначно визначити, оскільки вони будуть різні у кожного користувача, а дехто з користувачів матимуть декілька значень цих ознак. І якщо розглянуті вище системи-аналоги [6 – 16] працюють з виявленням аномалій по такій ознаці з невизначеною кількістю категорій-значень як IP-адреса, то по іншим таким ознакам вони не здійснюють виявлення шахрайства, що також впливає на їх ефективність, оскільки упускає важливі дані. До дихотомічних даних відносяться такі як: соціальні мережі (прив'язаний чи ні), інсталювання мобільного додатку (підтверджене чи ні), тип події (покупка чи ні), покупка (підтверджена чи ні).

Зазначимо, що більшість розглянутих вище аналогів [6 – 16] відкидають багато даних про користувачів. Проте на нашу думку використання усіх даних є важливим, тому що, як показано на рисунку 5, сукупність деяких даних може породжувати аномалію, яка є ознакою явного шахрайства. А відкинувши частину вхідних даних, існує можливість упустити деяких шахраїв та втратити інформацію для створення нових шаблонів і визначення основних їх ознак та характеристик. Так наприклад, на рисунку 5 за допомогою діаграм Венна подана множина вхідних даних A = {x_1, z_1, x_2, z_2, …, x_n, …, z_k, … | x ∈ X і z ∈ Z}, де елементи підмножини X = {x_1, x_2, … , x_n, …| $P_1(x)$ і $P_2(x)$ і … і $P_s(x)$} 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_k, … | $P_{a1}(z)$ або $P_{a2}(z)$ або … або $P_{k1}(z)$} у свою чергу не мають цих властивостей, що означає, що вони є аномальними у заданій множині даних A, та всі мають різні властивості – одну з: $P_{a1}(z)$ або $P_{a2}(z)$ або … або P_{k1}, тому вони хаотично розкидані, відповідно X ⊆ A та Z ⊆ A та X ⋂ Z = ∅.

Таким чином, різнорідність даних для задачі, що розглядається, настільки велика, що використовувати існуючі методи інтелектуального аналізу даних неможливо. Тому, виникла необхідність створення методів та алгоритмів, які б дозволили використовувати всю різнорідну інформацію про користувачів.

Метод подолання різнорідності даних. Проведений вище аналіз даних, а також досвід розробників систем-аналогів показав, що в купі всі ці дані неможливо використовувати, крім того, що вони різнорідні, їх дуже багато. І хоча сучасні методи інтелектуального аналізу даних можуть аналізувати величезну кількість даних, то вони не можуть подолати їх різнорідність (рисунок 4).

На нашу думку, для того, щоб використовувати всі наявні різнорідні дані, необхідно провести їх шкалювання як один із стандартних підходів подолання різнорідності. Але у зв'язку із складністю та кількістю даних, здійснимо шкалювання не однією, а декількома шкалами, об'єднуючи дані по групам. Такий підхід був запропонований авторами роботи у вигляді трьох алгоритмів подолання різнорідності на базі шести шкал згрупованих даних.

Алгоритм 1 подолання різнорідності вхідних даних при інсталюванні мобільних додатків:

1. Отримання даних $\overline{I}\begin{pmatrix} u_1(i_1, i_2, \ldots, i_{s1}) \\ u_2(i_1, i_2, \ldots, i_{s2}) \\ \ldots \\ u_n(i_1, i_2, \ldots, i_{sn}) \end{pmatrix}$.

2. Визначення типу даних.

3. Перетворення даних алгоритмами 2 та 3, в залежності від їх типу, з метою зведення їх до однорідних даних.

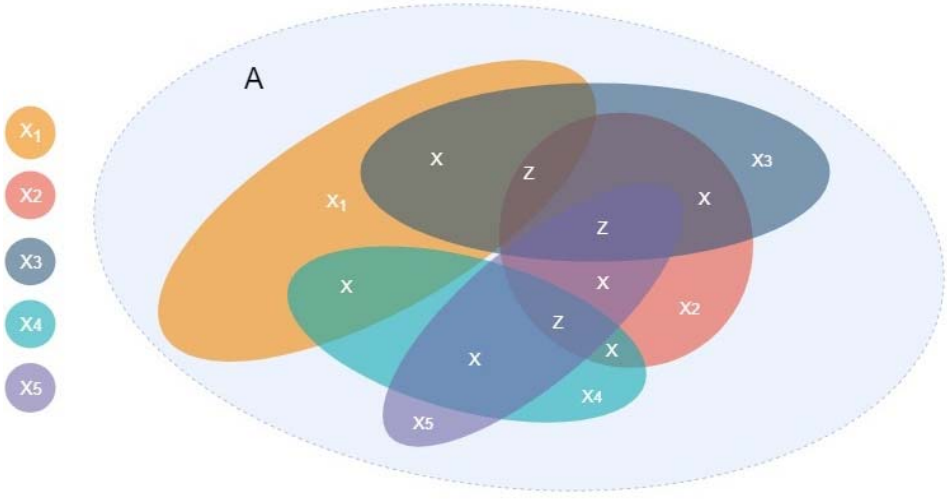

Рис. 5. Діаграма Венна, що зображає приклади шахрайства як навмисної аномалії в даних, де X_i – підмножини неаномальних даних, Z – підмножина аномальних даних

Для виконання пункту 3 алгоритму 1, пропонуємо алгоритм 2 шкалювання даних, який дає можливість переходу від різнорідних даних до коефіцієнтів, значення яких можуть бути або 0, що означатиме, що користувач є шахраєм, або 1, що позначатиме користувача як органічного. Тобто результатом шкалювання всіх даних являються 8 коефіцієнтів, значення яких бінарні – а саме 0 або 1. Бінарність значень коефіцієнтів визначається кінцевою ціллю задачі – чи користувач з певними даними являється шахраєм чи ні. Тому в першу групу G_1 входять дані, результатом аналізу яких є бінарна відповідь «так» або «ні» (0 або 1). А ті дані, які неможливо привести до бінарного типу, групуються у другу групу даних G_2, для яких у даній роботі розроблений метод (на базі алгоритму 3) інтелектуального аналізу даних з використанням коефіцієнтів схожості та системи правил з розробленою базою знань. Якщо ж даний метод не визначає користувача ні як шахрая, ні як органічного, то даний користувач включається в групу підозрілих користувачів, для аналізу якої застосовуються алгоритми нечіткої логіки на основі попередньо сформованих правил з бази знань, яка в процесі експлуатації постійно нарощується.

Слід зазначити, що не всі вхідні дані можна відразу ж однозначно шкалювати. Тому в роботі розроблений алгоритм 2 шкалювання вхідних даних, який оснований на розбитті всіх даних на 2 групи, над однією з яких однозначно проводиться шкалювання. Алгоритм 2 шкалювання вхідних даних:

1. Поділ усіх даних $\bar{I}$ на дві групи G_1 і G_2:

1.1. До першої групи G_1 входитимуть дані, за якими однозначно можна буде визначити коефіцієнт від 0 до 1, де 0 означатиме, що користувач є шахраєм, а 1 означатиме, що користувач є органічним (шкала 1, 2, 3 рисунок 6). Так, наприклад, значення коефіцієнту k_1, що відповідає визначенню шахрайства за IP-адресою, визначатиметься за допомогою перевірки належності IP-адреси користувача *IP* до множини відомих шахрайських IP-адрес *FRAUD_IP*, а саме, якщо виконується умова $P_1(IP) = IP \in FRAUD\_IP$, то користувач є шахраєм. Аналогічно, наприклад, значення коефіцієнту k_3, що відповідає визначенню шахрайства за унікальним ідентифікатором пристрою користувача *DeviceID*, визначатиметься умовою $P_3(DeviceID) = DeviceID \in FRAUD\_DeviceID$, виконання якої помічатиме користувача шахраєм. Зазначимо, що *FRAUD_DeviceID* – це множина усіх відомих ідентифікаторів пристроїв шахраїв. Коефіцієнт k_5 буде рівний значенню 0, якщо користувач відправив запит з ідентифікатором покупки *PurchaseID*, яка не була підтверджена мобільним магазином (атрибут покупки *isConfirmed* позначений як *false*). У цьому випадку визначення шахрая можна задати умовою $P_5(PurchaseID) = (PurchaseID.isConfirmed = false)$.

1.2. До другої групи G_2 входитимуть дані, за якими неможливо однозначно визначити значення коефіцієнту. Так, наприклад, невідомі граничні значення часу між подіями, за допомогою яких можна однозначно визначити, чи користувач є шахраєм чи органічним користувачем. Один із коефіцієнтів визначатиметься на основі типів подій, які робить користувач. Даний коефіцієнт не може входити до першої групи G_1, оскільки по таким даним не існує чітко визначеної умови, яка не буде змінюватися з часом.

2. Визначення значень коефіцієнтів на основі даних першої групи G_1 та побудова моделі даних групи G_1 (рисунок 6).

3. Визначення однозначних шахраїв, органічних та підозрілих користувачів на основі значень коефіцієнтів першої групи G_1, формування їх шаблонів та занесення їх у базу знань.

4. Визначення характеристик однозначно визначених користувачів, формування їх шаблонів та занесення їх у базу знань.

Для аналізу другої групи даних G_2 в роботі розроблено метод інтелектуального аналізу даних, основаного на алгоритмі 3:

1. Визначення значень коефіцієнтів другої групи даних G_2 на основі правил попередньо сформованої бази знань (шкали 4, 5 рисунок 7):

1.1. Отримання шаблонів та характеристик шахраїв з бази даних, що була сформована при аналізі групи G_1.

1.2. Знаходження коефіцієнту схожості між поточним користувачем та шаблоном і характеристиками шахрая (рисунок 7). Зазначимо, що значення коефіцієнтів схожості належить проміжку [0; 1]. Значення 1 означає, що користувачі ідентичні за даною ознакою, 0 у свою чергу означає, що користувачі не мають нічого спільного за даною ознакою. Так, наприклад, розглянувши такі дані як час між подіями, матимемо множину часу між подіями поточного користувача $T_U = \{t \mid t > 0\}$ та множини часу між подіями кожного із однозначно визначених на кроках алгоритму 3, 4 шахраїв $T_I = \{t \mid t > 0\}$. Маючи дві множини не бінарних, проте однорідних даних T_U та T_I, застосуємо відповідний коефіцієнт схожості користувачів. Для множин такого типу в роботі обрано коефіцієнт Танімото $K_T(T_U, T_I)$ [2, 4, 18], який визначається як $K_T(T_U, T_I) = \frac{N_C}{N_A + N_B - N_C}$, де N_C – кількість спільних для множин T_U та T_I елементів, N_A – кількість елементів у множині T_U, N_B – кількість елементів у множині T_I. У свою чергу, для множин бінарних даних, таких як множина з бінарними значеннями по кожному з існуючих типів подій мобільного додатку, де 0 означатиме, що користувач не використовував таку подію, а 1 означатиме,

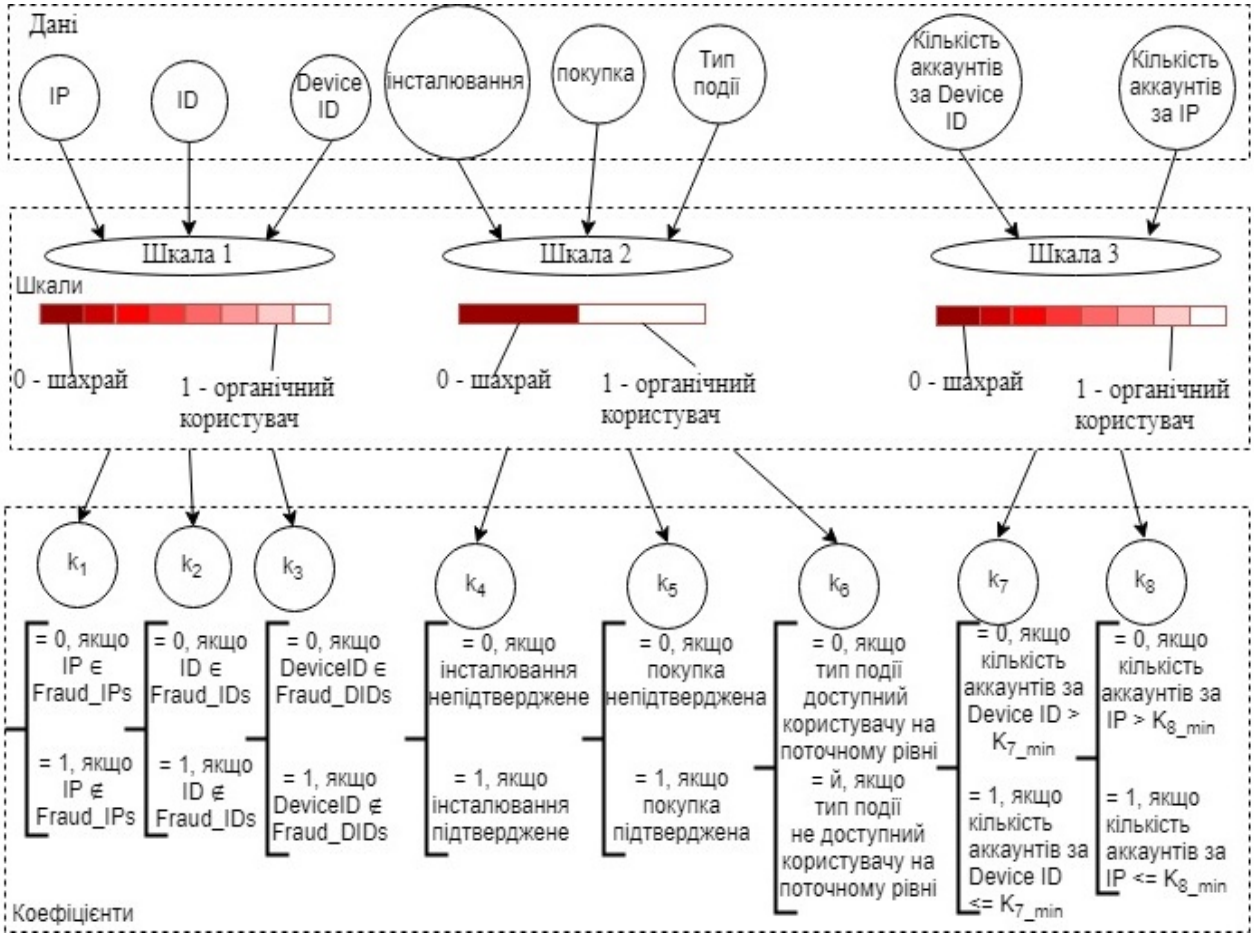

Рис. 6. Модель даних першої групи G_1

що використовував, коефіцієнти схожості користувачів визначаються з використанням коефіцієнту косинусної схожості $K_{COS}(A_1, A_2)$ [2, 4, 18], який найефективніше працює з бінарними даними. У свою чергу

$$K_{COS}(A_1, A_2) = \cos(A_1, A_2) = \frac{A_1 \cdot A_2}{|A_1| \cdot |A_2|},$$ де

A_1, A_2 – множини з вище вказаними бінарними даними поточного користувача та шахрайського користувача відповідно.

1.3. Формування масиву значень коефіцієнтів схожості поточного користувача з кожним із однозначно визначених шахрайських користувачів G_{2_fraud} з бази знань.

1.4. Виконання кроків 1.1 – 1.3 для однозначно визначених органічних користувачів G_{2_org} з бази знань.

1.5. Інверсія масиву значень коефіцієнтів схожості поточного користувача з кожним із однозначно визначених органічних користувачів G_{2_org} з бази знань.

1.6. Формування спільного масиву коефіцієнтів $G_2 = G_{2\_fraud} \cup (1 - G_{2\_org})$ по поточній ознаці на основі масивів, отриманих на кроках 1.3, 1.5, 1.6.

1.7. Формування вектору підозрілих користувачів G_{2_susp}.

2. Формування вектору відшкальованих коефіцієнтів по кожному користувачу, об'єднавши дані, отримані на кроці 2 алгориму 2 та кроці 1.7 алгоритму 3.

3. Віднесення підозрілих користувачів до класу шахраїв або органічних з використанням нечіткої логіки. Зв'язок між функціями належності входу i_j з бази знань можна визначати нечіткими логічними рівняннями

$$\mu^{d_j}(y) = b_{j1}\left[\mu^{j1}(i_1) \wedge \mu^{j1}(i_2) \wedge \ldots \wedge \mu^{j1}(i_n)\right] \vee$$

$$\vee\, b_{j2}\left[\mu^{j2}(i_1) \wedge \mu^{j2}(i_2) \wedge \ldots \wedge \mu^{j2}(i_n)\right] \vee \ldots$$

$$\vee\, b_{jp}\left[\mu^{jp}(i_1) \wedge \mu^{jp}(i_2) \wedge \ldots \wedge \mu^{jp}(i_n)\right], \; j = \overline{1, m},$$

які можна спростити до виразу

$$\mu^{d_j}(y) = \max_{p=1,k_j}\left\{a_{jp} \min_{i=1,n}\left[\mu^{jp}(i_j)\right]\right\}, \quad j = \overline{1, m},$$

де b_i^p – нечіткий терм. Нечіткий терм у свою чергу

визначається як $b_i^p = \int_{\underline{i_j}}^{\overline{i_j}} \frac{\mu^p(i_j)}{i_j}$, де $\mu^p(i_j)$ – функція належності входу i_j нечіткому терму b_i^p, $p = k_l$, $i = \overline{1,n}$, $j = \overline{1,m}$.

В результаті виконання кроків 1 та 2 алгоритму 2, проведено шкалювання даних групи G_1 та за допомогою коефіцієнтів k_1, k_2, k_3, …, k_8 різнорідні дані приведено до однорідного стану, що дозволяє однозначно визначити шахраїв, якщо значення хоча б одного з коефіцієнтів дорівнює 0.

Зазначимо, що запропоновані метод та алгоритми алгоритм можна автоматизувати. Також, з кожним повторним використанням алгоритмів, будуть знаходитись все нові характеристики органічних та шахрайських користувачів, які заноситимуться у базу знань. Це дозволить удосконалювати подальше виявлення шахрайства.

Таким чином, запропонований метод подолання різнорідності вхідних складається з трьох алгоритмів: алгоритм 1 подолання різнорідності даних, алгоритм 2 шкалювання даних, алгоритм 3 визначення коефіцієнтів подібності з використанням інтелектуального аналізу даних на основі нечіткої логіки, та двох розроблених моделей даних, що дозволяє суттєво підвищити точність прийняття рішення та точність кінцевих результатів, що підтверджено результатами експериментальних досліджень.

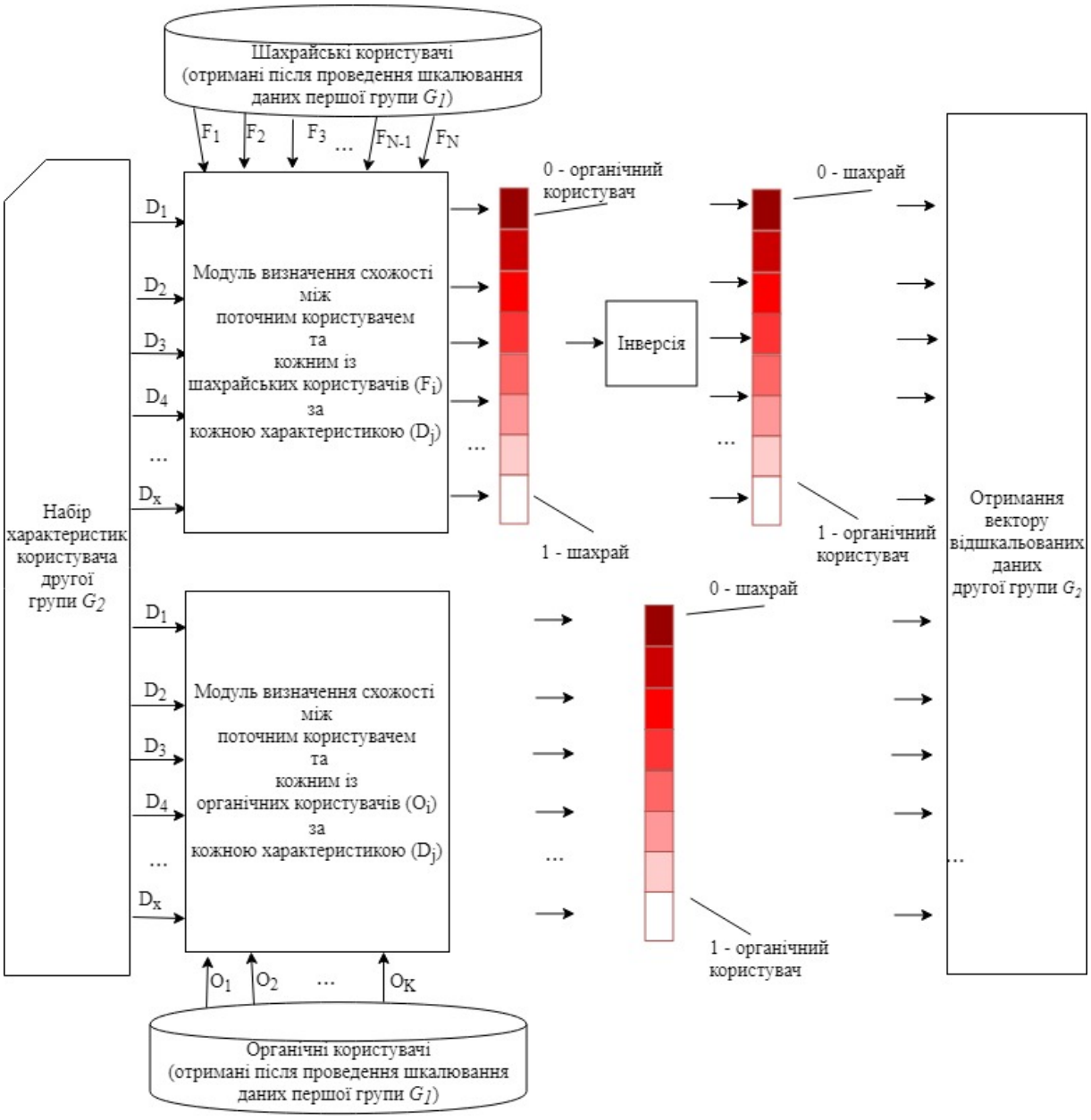

Рис. 7. Модель шкалювання даних другої групи G_2

Результати експериментальних досліджень. У роботі розроблено програмне забезпечення, в основі якого лежать запропоновані алгоритми подолання різнорідності. Для проведення експерементальних досліджень взято вибірку з 300567 записами з реального мобільного додатку, всі дані якої різнорідні та кожен користувач якої помічений класом (шахрай чи органічний). Розроблена авторами система не знає про мітки користувачів, а помічений набір було обрано для подальшої перевірки точності розробленої системи. Точність виявлення шахрайства при інсталюванні мобільних додатків на обраній вибірці склала 99,14 % в результаті того, що подолана різнорідність даних за допомогою запропонованого вище методу та алгоритмів, що дало можливість використовувати усі дані про користувачів, та використано метод виявлення схожості користувачів [17 – 18]. Також, завдяки розробленому методу подолання різнорідності даних, були більш точно визначені характеристики шахраїв та сформовані правила в базу знань. Для реалізації обрано мову програмування Python та бібліотеку TensorFlow, розроблену компанією Google. Результати експериментального дослідження, що розроблені на основі програмних модулів, на які отримано авторські свідоцтва [19 – 20] представлено на рисунку 8.

```
Percent of fraudulent transactions:  0.0017274856306200 34
/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. 8. Результати комп'ютерного моделювання, здійсненого на основі даних, отриманих з методу подолання різнорідності вхідних даних

Висновки. Наукову новизну запропонованої роботи складають:

Розроблено метод подолання різнорідності даних при виявленні шахрайства під час інсталювання мобільних додатків, який на відміну від існуючих, використовує усю множину вхідних даних, що дозволяє знаходити нові шахрайські шаблони, їх характеристики та залежності та підвищує ефективність процесу виявлення шахрайства при інсталюванні мобільних додатків.

Розроблений метод складається з трьох запропонованих алгоритмів: алгоритм 1 подолання різнорідності вхідних даних, алгоритм 2 шкалювання даних, алгоритм 3 визначення коефіцієнтів подібності з використанням інтелектуального аналізу даних на основі нечіткої логіки, та двох розроблених моделей даних. Таке поєднання алгоритмів дозволяє суттєво підвищити точність прийняття рішення та точність кінцевих результатів.

1. Результати експериментальних досліджень на вибірці з 300567 записів з реального мобільного додатку, всі дані якої різнорідні та кожен користувач якої помічений класом (шахрай чи органічний), дозволили отримати точність 99,14 %.

Практична цінність підходу заключається у можливості його використання при вирішенні задач автоматичного виявлення шахрайства при інсталюванні мобільних додатків за допомогою спеціально розробленого програмного забезпечення.

Література

1. *S. Benndorf, G. Kakulapati, A. Pham and others* (2015), "Fighting Mobile Fraud in the Programmatic era: AppLift". AppLift GmbH, 14 p.
2. *А. А. Яровий, О. Н. Романюк, І. Р. Арсенюк, та Т. Д. Польгуль* (2017) "Виявлення шахрайства при інсталюванні програмних додатків з використанням інтелектуального аналізу даних". Наукові праці Донецького національного технічного університету Серія: "Інформатика, кібернетика та обчислювальна техніка", Покровськ, №2 (25) – С. 126-131.
3. **"**Our take on mobile fraud detection", available at: http://geeks.jampp.com/data-science/mobile-fraud/
4. *Vacha Dave, Saikat Guha, Yin Zhang,* "ViceROI: Catching Click-Spam in Search Ad Networks", available at: http://www.sysnet.ucsd.edu/~vacha/ccs13.pdf
5. *Dave, V., Guha, S., Zhang Y.* (2012), "Measuring and Fingerprinting Click-Spam in Ad Networks. In Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication (SIGCOMM)". Helsinki, Finland, Aug. 2012. – 175 – 186 pp.
6. "Кракен Антибот", available at: http://kraken.run/
7. "Fraudlogix: Ad Fraud Solutions for Exchanges, Networks, SSPs & DSPs", available at: https://www.fraudlogix.com/
8. "AppsFlyer: Measure In-App To Grow Your Mobile Business", available at: https://www.appsflyer.com/
9. "Adjust", available at: https://www.adjust.com/
10. "Kochava Uncovers Global Ad Fraud Scam", available at: https://www.kochava.com/
11. "TMC Attribution Analytics", available at: https://help.tune.com/marketing-console/attribution-analytics/
12. **"**Fraudwatch", available at: http://www.fraudshields.com/
13. **"**Impact: Forensiq by Impact Earns MRC Accreditation for SIVT Detection and Filtration and Viewability Measurement", available at: https://impact.com/ad-fraud-detection/
14. "Appsflyer: Protect your data from mobile fraud: Protect360", available at: https://www.appsflyer.com/ product/ protect360/
15. "FraudScore: FraudScore fights ad fraud using Machine Learning", available at: https://fraudscore.mobi/
16. "AppMetrica: Аналитика приложений от и до", available at: https://appmetrica.yandex.ru/
17. *Aurélien Géron* (2017), "Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and

Techniques to Build Intelligent Systems". O'Reilly Media, 574 p.
18. *Кюльян А. Г., Польгуль Т. Д., Хазін М.Б.* (2012), "Математична модель рекомендаційного сервісу на основі методу колаборативної фільтрації", Комп'ютерні технології та Інтернет в інформаційному суспільстві, 226–227 с.
19. *А.А. Яровий, Т.Д. Польгуль.* (2018). 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, 26.01.18.
20. *А.А. Яровий, Т.Д. Польгуль.* (2018). 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, 26.01.18.

References

1. *S. Benndorf, G. Kakulapati, A. Pham and others* (2015), "Fighting Mobile Fraud in the Programmatic era: AppLift". AppLift GmbH, 14 p.
2. *A. A. Yarovyi, O. N. Romanyuk, I. R. Arsenyuk, and T. D. Polhul* (2017) " Application Install Fraud Detection Using Data Mining ". *Scientific Works of Donetsk National Technical University, Series: "Informatics, cybernetics and computing technique"*, Pokrovsk, №2 (25) – p. 126-131.
3. **"**Our take on mobile fraud detection", available at: http://geeks.jampp.com/data-science/mobile-fraud/
4. *Vacha Dave, Saikat Guha, Yin Zhang,* "ViceROI: Catching Click-Spam in Search Ad Networks", available at: http://www.sysnet.ucsd.edu/~vacha/ccs13.pdf
5. *Dave, V., Guha, S., Zhang Y.* (2012), "Measuring and Fingerprinting Click-Spam in Ad Networks. In Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication (SIGCOMM)". Helsinki, Finland, Aug. 2012. – 175 – 186 pp.
6. Kraken Antibot", available at: http://kraken.run/
7. Fraudlogix: Ad Fraud Solutions for Exchanges, Networks, SSPs & DSPs", available at: https://www.fraudlogix.com/
8. "AppsFlyer: Measure In-App To Grow Your Mobile Business", available at: https://www.appsflyer.com/
9. "Adjust", available at: https://www.adjust.com/
10. "Kochava Uncovers Global Ad Fraud Scam", available at: https://www.kochava.com/
11. "TMC Attribution Analytics", available at: https://help.tune.com/marketing-console/attribution-analytics/
12. **"**Fraudwatch", available at: http://www.fraudshields.com
13. **"**Impact: Forensiq by Impact Earns MRC Accreditation for SIVT Detection and Filtration and Viewability Measurement", available at: https://impact.com/ad-fraud-detection/
14. "Appsflyer: Protect your data from mobile fraud: Protect360", available at: https://www.appsflyer.com/ product/protect360/
15. FraudScore: FraudScore fights ad fraud using Machine Learning", available at: https://fraudscore.mobi/
16. "AppMetrica", available at: https://appmetrica.yandex.ru/
17. *Aurélien Géron* (2017), "Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems". O'Reilly Media, 574 p.
18. *Kiulian A. H., Polhul T. D., Khazin M.B.* (2012), "Mathematical model of recommendation service based on the method of collaborative filtration" (Ukr.), Computer Technologies and Internet in information society, 226–227 pp.
19. *A.A. Yarovyi, T.D. Polhul*, Computer Program «Software module for data gathering of information technology of fraud detection in the process of software applications installation», Certificate of registration of copyright to a work №76348, K.: Ministry of Economic Development and Trade of Ukraine, 26.01.18. (Ukr.).
20. *A.A. Yarovyi, T.D. Polhul*, Computer Program «Software module of users similarity definition of information technology of fraud detection in the process of software applications installation», Certificate of registration of copyright to a work №76347, K.: Ministry of Economic Development and Trade of Ukraine, 26.01.18. (Ukr.).

Польгуль Т. Д., Яровой А. А. Метод преодолений разнородности данных для выявления мошенничества при инсталлировании мобильных приложений

В работе предложено метод и алгоритмы преодоления разнородности данных для выявления мошенничества при инсталлирования мобильных приложений. Процедура выявления мошенничества на основе разработанного метода позволяет выявить мошенников и определить их характеристики и шаблоны. Осуществлены экспериментальные исследования на основе выбранной помеченной выборки. В экспериментах метки классов не использовались, однако они необходимы для проверки точности процедуры принятия решений, разработанной на основе предложенного метода, которая составила 99,14%.

Ключевые слова: *выявление мошенничества, преодоление разнородности, интеллектуальный анализ данных, выявление аномалий, инсталлирование мобильных приложений.*

Polhul T., Yarovyi A. The input data heterogeneities resolution method during mobile applications installation fraud detection

The data heterogeneities resolution method and algorithms during mobile applications installation fraud detection were proposed in the paper. The fraud detection procedure on the basis of the developed method allows to detect scammers, their characteristics and patterns. The experimental studies on the basis of the selected labeled sample were performed. Samples class labels were not used in experiments, but they are necessary to verify the accuracy of the decision-making procedure developed on the basis of the proposed method system, which is 99,14 %.

Keywords: *fraud detection, heterogeneities resolution, data mining, anomaly detection, mobile applications installation.*

Польгуль Т. Д.– аспірант кафедри комп'ютерних наук Вінницького національного технічного університету, e-mail: tanapolg93@gmail.com
Яровий А. А. – д.т.н., професор, завідувач кафедри комп'ютерних наук Вінницького національного технічного університету, e-mail: a.yarovyy@vntu.edu.ua

Рецензент: д.т.н., проф. ***Рязанцев О.І.***

Стаття подана 16.11.2018.