## ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

**УДК 004.8:044.89**

Т. Д. Польгуль

# ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ПОБУДОВИ ІНТЕЛЕКТУАЛЬНИХ СИСТЕМ ВИЯВЛЕННЯ ШАХРАЙСТВА ПРИ ІНСТАЛЮВАННІ МОБІЛЬНИХ ДОДАТКІВ

Вінницький національний технічний університет, Вінниця

Анотація. У роботі запропоновано інформаційну технологію побудови інтелектуальних систем виявлення шахрайства при інсталюванні мобільних додатків, яку бажано використовувати при розробці такого класу систем. Здійснено інтелектуальну обробку наявних даних по користувачу, на основі якої запропоновано шкалювання не по значенню ознаки, а по кінцевій інформації, яку несе ознака по користувачу. Запропоновано систему з інтелектуальною складовою – формуванням бази знань, яка дозволить визначити шахраїв, та в яку включатимуться правила аналізу аномалій, при чому так, щоб поява нової аномалії в даних дозволяла створити нове правило. Така база знань може розширюватись через можливість появи нового виду аномалії в даних (шахрайства). Отриманий набір правил, що у подальшому на основі розроблених в роботі алгоритмів дозволить створити узагальнений портрет шахрая, відзначивши навіть нові та невідомі експертам шахрайські властивості. Віднесення підозрілих користувачів до класу шахраїв або органічних відбувається з використанням нечіткої логіки. Завдяки запропонованій інтелектуальній обробці наявних даних про користувачів, шкалюванню по кінцевій інформації, яку несе ознака та створенню баз знань, що розвиваються, запропонована інформаційна технологія дозволяє побудувати інтелектуальні системи, що матимуть можливість адаптуватися до появи нових видів шахрайства. Відповідно до задач, які повинні вирішувати такі інтелектуальні системи, запропоновано їх структуру: підсистема виявлення характеристик користувача; підсистема подолання різнорідності даних; підсистема тренування класифікаційної моделі; підсистема класифікації; підсистема формування бази даних шахраїв; підсистема формування бази знань (для виявлення шахраїв); підсистема інтелектуального аналізу даних та формування шаблонів користувачів; підсистема прогнозування узагальненого шаблону шахрая. Запропонована інформаційна технологія побудови інтелектуальних систем дозволяє обробляти різноформатні вхідні дані, що у процесі дає можливість сформувати узагальнений шаблон шахрая.

**Ключові слова**: **виявлення шахрайства, виявлення аномалій, інсталювання мобільних додатків, інтелектуальний аналіз даних, інформаційна технологія, інтелектуальні системи.**

Аннотация. В работе предложена информационная технология построения интеллектуальных систем обнаружения мошенничества при инсталлирования мобильных приложений, которую желательно использовать при разработке такого класса систем. Осуществлено интеллектуальную обработку имеющихся данных по пользователю, на основе которой предложено шкалирование не по значению признака, а по конечной информации, которую несет признак по пользователю. Предложена система с интеллектуальной составляющей – формированием базы знаний, которая позволит определить мошенников, и в которую будут включаться правила анализа аномалий, причем так, чтобы появление новой аномалии в данных позволяло создать новое правило. Такая база знаний может расширяться из-за возможности появления нового вида аномалии в данных (мошенничества). Полученный набор правил, в дальнейшем на основе разработанных в работе алгоритмов позволит создать обобщенный портрет мошенника, отметив даже новые и неизвестные экспертам мошеннические свойства. Отнесение подозрительных пользователей к классу мошенников или органических происходит с использованием нечеткой логики. Благодаря предложенной интеллектуальной обработке имеющихся данных о пользователях, шкалирования по конечной информации, которую несет признак и созданию баз знаний развивающихся предложена информационная технология позволяет построить интеллектуальные системы, которые будут иметь возможность адаптироваться к появлению новых видов мошенничества. В соответствии с задачами, которые должны решать такие интеллектуальные системы, предложено их структуру: подсистема обнаружения характеристик пользователя; подсистема преодоления разнородности данных; подсистема тренировки классификационной модели; подсистема классификации; подсистема формирования базы данных мошенников; подсистема формирования базы знаний (для выявления мошенников); подсистема интеллектуального анализа данных и формирования шаблонов пользователей; подсистема прогнозирования обобщенного шаблона мошенника. Предложенная информационная технология построения интеллектуальных систем позволяет обрабатывать разноформатные входные данные, что в процессе дает возможность сформировать обобщенный шаблон мошенника.

**Ключевые слова: определение мошенничества, определение аномалий, инсталлирование мобильных приложений, интелектуальный анализ данных, информационная технология, интелектуальные системы.**

Abstract. Information technology for the construction of intelligent systems for detecting fraud during mobile applications installations, which is desirable to use in developing such a class of systems, is proposed in this paper. The intelligent processing of available data by the user is done. The scaling which is based not on the value of the feature, but on the end-point information of the feature, is proposed based on this intelligent processing of the data. A system with an intellectual component - the formation of a knowledge base that will allow fraudsters to be identified and which will include anomaly analysis rules - was proposed, so that the emergence of a new anomaly in the data allows for the creation of a new rule. Such knowledge base can be expanded due to the possibility of an emergence of a new kind of anomaly in data (fraud). The received set of rules will allow creating a generalized fraudster's fingerprint, noting even the new and unknown for experts fraudulent patterns, based on the algorithms developed in the work. The classification of suspicious users to a class of fraudsters or organic users is made using fuzzy logic. The information technology for the construction of intelligent systems that will be able to adapt to the emergence of new types of fraud is proposed based on the proposed intelligent processing of available user data, the scaling by end-point information, and the development of knowledge bases. According to the tasks which should be solved by such intelligent systems, their structure is proposed: subsystem of user data characteristics identifying; subsystem of overcoming heterogeneity; subsystem of classification model training; subsystem of classification; subsystem of fraudsters database formation; subsystem of knowledge base (for detecting fraud) formation; subsystem of data mining and user patterns formation; subsystem of general fraudster fingerprint prediction. The proposed information technology for the construction of intelligent systems allows processing of various input data, which in the process gives the opportunity to form a generalized fraudster fingerprint.

**Keywords: fraud detection, anomaly detection, mobile application installation, data mining, information technology, intellectual systems.**

**DOI: https://doi.org/10.31649/1999-9941-2019-44-1-4-16.**

## Вступ

У зв'язку з появою на ринку величезної кількості мобільних додатків, якими користуються мільярди користувачів, компанії-розробники мобільних додатків користуються послугами маркетингових кампаній з метою залучення користувачів саме до їхнього додатку. Саме така потреба у маркетингових кампаніях стала однією з причин появи шахраїв та їх шахрайських способів інсталювання мобільних додатків. Шахраї у свою чергу приводять компаніям необхідну кількість «користувачів» та отримують за це відповідну грошову винагороду, проте їхні «користувачі» ніколи не повертаються у мобільний додаток, оскільки є фейковими, ми ж їх називатимемо шахрайськими.

На наш час вже існують такі відомі види шахрайства при інсталюванні додатків, як мобільне викрадення (mobile hijacking), кліковий спам (click spamming), ферми дій (action farms) [1-2], а також методи та системи виявлення шахрайства при інсталюванні мобільних додатків такі як Fraudlogix та Kraken, Adjust, Kochava та TCM Attribution Analytics, Protect360 від Appsflyer, FraudScore та AppMetrica, що згадані у роботі [3]. Проте необхідно зазначити, що лише остання пара використовує інтелектуальну складову, серед них AppMetrica просто опирається на FraudScore та використовує їх алгоритми та API, але навіть вказані системи-аналоги виконують рейтингування користувачів на основі не всіх, а вибіркових вхідних даних, тому наявне упущення шахраїв системою [4]. Інші вказані системи використовують відомі бази з шахрайськими даними (наприклад, IP-адресами), що також призводить до упущення шахраїв, що мають інші властивості, шаблони, поведінку.

Очевидно, що причиною вищевказаних недоліків системи є відсутність єдиної концепції виявлення шахрайства на основі всіх наявних даних. Також, недоліком існуючих систем є те, що вони розпізнають лише відомі види шахрайства і не можуть розпізнавати нові шахрайські шаблони. А в сучасному світі важливою є можливість системи адаптуватись, тому необхідним є створення інтелектуальної системи, що матиме змогу самонавчатися.

Тому й виникає необхідність створення інформаційної технології побудови інтелектуальних систем виявлення шахрайства при інсталюванні мобільних додатків, яка б відстежувала та визначала шаблони шахраїв, що непомітні людині. Для вирішення поставленої мети в даній роботі запропонована інформаційна технологія побудови інтелектуальних систем виявлення шахрайства при інсталюванні мобільних додатків.

## Мета

Метою статті є створення інформаційної технології побудови інтелектуальних систем виявлення шахрайства при інсталюванні мобільних додатків, особливістю якої була б можливість адаптуватись до появи нових видів шахрайства.

## Задачі дослідження

У процесі розробки інформаційної технології побудови інтелектуальних систем виявлення шахрайства при інсталюванні мобільних додатків виникли наступні задачі:

– виявлення, аналіз та використання наявних даних різних шаблонів, розмірності, метрик методами класифікації та шкалювання;

– інтелектуальна обробка даних про користувачів методами класифікації, коефіцієнтів схожості та нечіткої логіки;

– класифікація користувачів з використанням моделі глибоких нейронних мереж;

– створення баз даних, що містять характеристики шахраїв не лише людини, а і різних програмних ботів, та баз знань з набором нечітких правил визначення шахраїв. При чому такі бази даних та бази знань повинні розвиватися в залежності від нових даних та нових шахраїв з метою створення узагальненого портрету шахрая.

Розв'язання вище перерахованих задач потребує розробки нових моделей, методів, алгоритмів та засобів і програмного забезпечення, що дозволить створити інформаційну технологію побудови інтелектуальних систем виявлення шахрайства при інсталюванні мобільних додатків, включаючи розробку структури, методів обробки даних та алгоритмів функціонування таких систем.

Розглянемо вирішення кожної задачі для побудови інформаційної технології побудови інтелектуальних систем виявлення шахрайства при інсталюванні мобільних додатків та проаналізуємо її ефективність на тестовій системі.

## Визначення поняття шахрайства

Спочатку розглянемо, в чому полягає особливість систем виявлення шахрайства. Зазвичай результатом роботи таких систем є таблиця рейтингування користувачів, яка в процентному відношенні показує, відноситься користувач до шахрая чи ні, так наприклад працює система-аналог FraudScore [5]. Але як оцінювати користувача, який в такій таблиці має відсоток 49, чи є він шахраєм чи ні, чи система підозрює його? Тобто така таблиця дає достатньо нечітку інформацію, за результатом якої складно приймати

рішення, визначати такого користувача шахраєм чи ні. Тому для початку необхідно дати визначення поняття «шахрайство».

На наш погляд шахрайство можна розглядати як аномалію в даних. Розглянемо відоме поняття «аномалія в даних» у сфері інформаційних технологій, визначення якого представлені у роботі [6]. Так, вчені з університету Мінессота визначають у [7], що аномалія – це шаблон даних, який не відповідає визначеному поняттю нормальної поведінки (заданому шаблону). При цьому аномалії можуть бути наявними у вхідних наборах даних систем прийняття рішень та систем штучного інтелекту через неуважність персоналу, який вносив ці дані, через наявність похибок та некоректну роботу системи збору даних або ж через навмисні дії шахраїв (іншими словами – шахрайство). Наявність аномалій в даних приводить до помилок в процесі прийняття рішень, так наприклад:

– неправильне визначення діагнозу у медичних системах прийняття рішень та системах штучного інтелекту у медичній сфері, що може коштувати пацієнту життя;

– невизначення механічних несправностей у літаках, машинах тощо перед їх експлуатацією, що також може коштувати життя людям;

– неправильне визначення або невизначення шахрайства у банківських транзакціях. У випадку крадіжки картки та виконання незвичних транзакцій з неї, неправильно навчена система не заблокує шахрайську транзакцію.

Поняття аномалії згадується в літературі з аналізу даних (data mining) як нетипова поведінка, аномальність, викид (outliers), відхилення. Про це також зазначають вчені з Тернопільського національного економічного університету [8]. Існують визначення, які вважаються досить загальними. Так наприклад Хокінз (Hawkins, 1980) у [9] визначає викид (аномалію) як спостереження, яке відхиляється від інших спостережень настільки, що виникають підозри, що він був породжений іншим механізмом. Барнет і Льюїс (Barnet and Lewis, 1994) вказують у [10] на те, що викид (аномалія) є тим, що помітно відхиляється від інших зразків, в яких воно виявляється. Аналогічно, Джонсон (Johnson, 1992) визначає викид (аномалію) як спостереження у наборі даних, яке суперечить іншій частині цього набору даних. Але в даних прикладах не дається чітке визначення поняття аномальності, яке може бути використане для розробки математичних моделей та програмного забезпечення. З іншого боку, простий приклад аномалії у двовимірному просторі представлено на рисунку 1. На ньому показано, що найпростіший приклад аномалії в даних (множина $A$) – це відокремлена підмножина неаномальних даних $X$, яка містить свої властивості та підмножина аномальних даних $Z$, яка характеризується своїми властивостями.

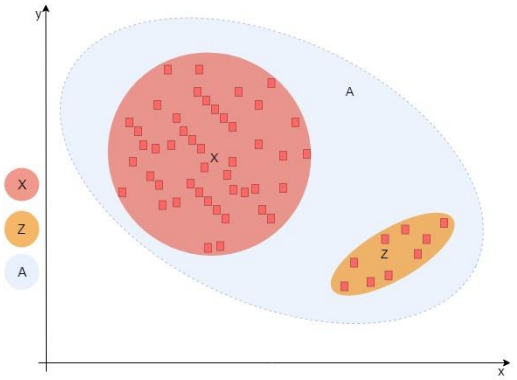



Рисунок 1 – Простий приклад аномалії у двовимірному просторі, де $X$ – підмножина неаномальних даних, $Z$ – підмножина аномальних даних, $A$ – множина усіх вхідних даних

У даній роботі аномалію в даних розглядатимемо як групу (множину) даних $\langle a_1, a_2, ...a_n \rangle$, яка входить у множину вхідних даних $A$, що характеризується множиною властивостей $P(a)$, але виходить за задані межі цих властивостей. У задачі виявлення аномалій у вхідних наборах даних з мобільних додатків, аномальними будемо вважати ті дані (елементи множини), які:

– не мають властивостей $P_1(x)$, $P_2(x)$, …, $P_S(x)$, що визначають множину неаномальних даних $X$;

– не входять в область гранично допустимих значень множини неаномальних даних $X$ – задамо це властивістю $P2(x)$, яка має вигляд ( $x <= \max\_value$ і $x >= \min\_value$ );

– не співпадають по розмірності;

– не співпадають по властивостям групи даних.

Шахрайство ж у свою чергу визначимо як навмисне породження аномалії в даних сторонньою особою (шахраєм) або механізмом з певною метою. Аномалії, як і шахрайство, можуть бути наявні у всіх наборах даних, а отже й у всіх областях, з якими працюють системи прийняття рішень та системи штучного інтелекту. Тому на наш погляд більш вдалим означенням шахраїв при розробці такого класу систем – це визначати наявність шахраїв як наявність аномалій в масивах вхідних даних. Проте як же визначати аномалії в даних? Спочатку розглянемо дані, які використовуються в системі.

**Аналіз вхідних даних**

Детальний аналіз даних проведений в роботі [6] показав, що дані в таких системах є різнорідними, а саме – різних шаблонів, метрик, розмірностей. На рис. 2 представлені всі дані, які використовуються в такому класі систем. Дані згруповані відносно груп подій, на рис. 3 всі наявні для такого класу систем дані зведено до вигляду таблиці і показано, що ці дані різняться не лише за метриками, розмірностями і шаблонами, але також є як якісні, кількісні дані, так і масиви якісних і кількісних даних, тому зрозуміло, що аналіз даних – це дуже складна задача.

Під різнорідністю розуміється, шо усі дані, а саме дані з різних класів, неможливо порівняти між собою. Також слід зазначити, що серед наявних даних таких систем є не лише якісні і кількісні дані, але й масиви якісних та кількісних даних. Давайте поглянемо, які дані є (рис. 2, 3).

З рис. 3 також видно, що всі дані є різнорідними, а так як працювати з такими даними складно, багато існуючих систем, методів та моделей відкидають, наприклад, масиви якісних даних чи дані, всі значення яких завчасно невідомі тощо. Так наприклад, кожен з користувачів мобільного додатку має такий якісний параметр як IP-адресу – це і є однією з характеристик, якою часто нехтують. Зазначимо, що деякі з систем перевіряють наявність поточної IP-адреси у відомій базі даних з шахрайськими IP-адресами, але не подають її у якості ознаки (feature) в систему інтелектуального аналізу даних. Це відбувається через ряд причин: моделі інтелектуального аналізу даних в основному працюють лише з числовими значеннями; для перетворення якісних даних у кількісні зазвичай використовується метод one-hot encoding [11], але для його використання необхідна завчасно відома множина можливих значень ознак (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 версій протоколів IPv4 та IPv6, проте це призведе до зберігання величезної кількості надлишкової інформації, що значно погіршить швидкодію та ефективність роботи системи. Тому зазвичай такі дані відкидаються і не приймають рішення у процесі прийняття кінцевого рішення – користувач шахрай чи ні. Але ж чим більша кількість вхідних даних, тим більше залежностей, кореляцій можна з них виділити та тим більше різних шаблонів шахраїв можна буде помітити. Тобто чим більше даних для прийняття рішень, тим вища точність системи. Тому перед авторами постало питання як правильно та ефективно використовувати усі різнорідні дані про користувача, оскільки практично всі системи класифікації працюють з однорідними даними.

Зазвичай в різних областях науки і техніки для аналізу таких даних використовують різні методи шкалювання, так наприклад:

– за допомогою систем правил, побудованих з використанням нечіткої логіки;

– нормалізацією даних, що використовується наприклад у задачах статистичної обробки даних;

– звичайним шкалювання вимірювальних пристроїв.

Проте, як можна для даних про користувачів при інсталюванні мобільних додатків, що представлені на рис. 2, 3, провести шкалювання? Відомо, що IP-адреса може мати вигляд наприклад такий як «127.0.0.1», «192.168.5.1». Нормалізувати дані такого типу неможливо. Аналогічно недоцільно використовувати систему правил для даних такого типу. Оскільки для того, щоб взяти до уваги всі ці дані, необхідно сформувати 4–12 правил, що суттєво ускладнює процес аналізу. Отже, серед вказаних трьох варіантів залишається шкалювання. Проте на думку авторів при аналізі таких даних традиційне шкалювання всього набору даних не є можливим, оскільки значення у всіх цих даних різне (число, категорія якісних даних, масиви кількісних чи якісних даних), а інформація, яку вони несуть, однозначно визначає чи дана ознака характеризує шахрая чи має властивості органічного користувача. Оскільки важливим є здійснити не лише рейтингування користувача, але й зазначити причину, чому даного користувача помітили шахраєм, що є важливим при розгляді судових позовів, в яких необхідно чітко наводити аргументи. Так наприклад, це одна з найвагоміших причин, чому безпілотні автомобілі ще не доступні у продажу, адже в досліджуваній області таке обґрунтування є обов'язковим [6]. Тому в роботі [3] був запропонований метод шкалювання даних про користувача по цінності інформації, яку вони несуть.

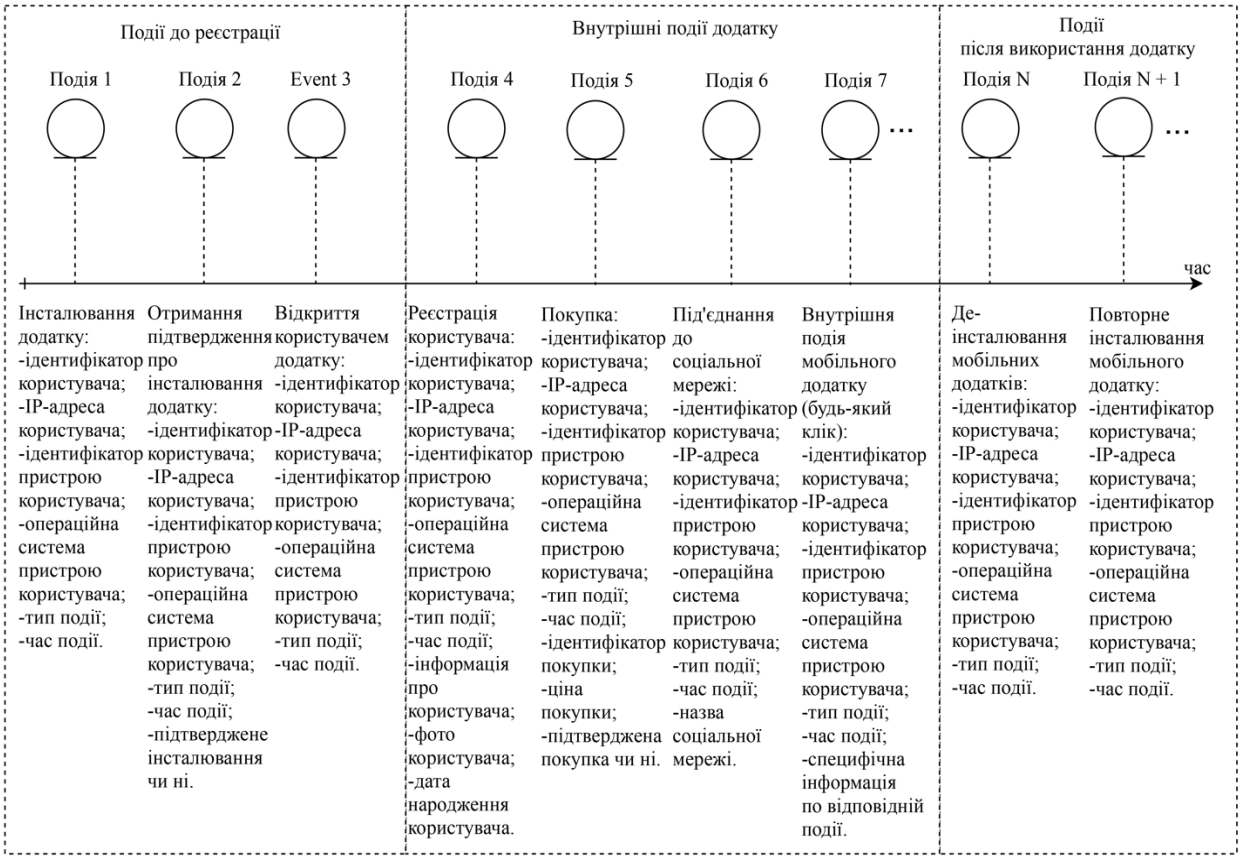



Рисунок 2 – Приклад вхідних даних про користувача в системах виявлення шахрайства при інсталюванні мобільних додатків

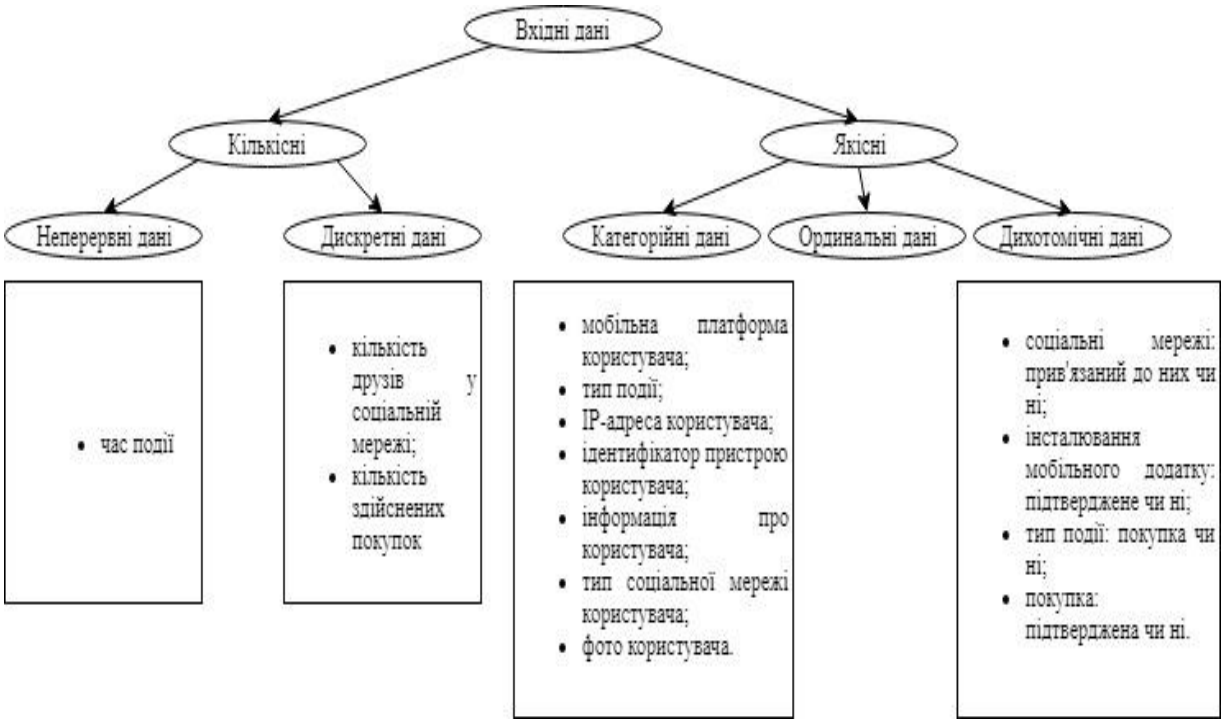



Рисунок 3 – Класифікація різнорідних вхідних даних в системах виявлення шахрайства при інсталюванні мобільних додатків

Отже, для подолання різнорідності даних використовується шкалювання по цінності інформації, яку несуть ці дані, що означає перетворення всіх даних (якісних і кількісних) до єдиної бінарної шкали зі значеннями 0 та 1 в залежності не від значення, а від інформації, яку несуть ці дані. Наприклад, розглянемо шкалу для переведення ознаки «IP-адреса» до бінарного значення, представлену на рис. 4. Після шкалювання за даною ознакою, деяких користувачів можна однозначно віднести до шахраїв. Вважатимемо, що значення 0 на шкалі означатиме, що користувач є шахраєм за даною ознакою, а значення 1 на шкалі означатиме, що користувач є органічним (тобто не є шахраєм) за даною ознакою. Це є особливістю запропонованого методу шкалювання різнорідних даних.

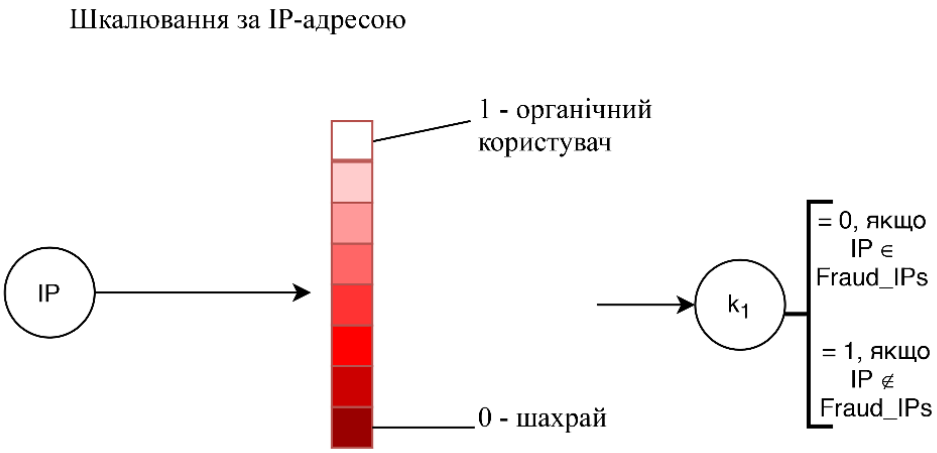



Рисунок 4 – Шкалювання за IP-адресою

Однак, при використанні такого шкалювання не всіх користувачів можна однозначно віднести або до шахраїв, або до органічних, оскільки ця IP-адреса може буде відсутня в базі даних, тому для таких користувачів необхідно додатково використати інтелектуальну обробку даних.

**Інтелектуальна обробка наявних даних про користувачів**

В залежності від існуючих даних і при використанні запропонованого методу шкалювання в роботі було розроблено 17 шкал, особливістю кожної шкали є те, що вони бінарні. Кількість шкал визначалася як кількістю вхідних даних, так і результатами експертного оцінювання, розглянутого у роботі [6], на основі якого розроблений метод подолання різнорідності даних в системах, що розглядаються. Шкали у свою чергу відповідають конкретній інформації про користувача, так наприклад, важливими є IP-адреса користувача, покупки користувача, інформація про під'єднання користувача до соціальних мереж, час інсталювання мобільного додатку користувачем, час між внутрішніми подіями додатку у користувача тощо.

Для того, щоб кожну з даних ознак представити у бінарному вигляді, було сформовано коефіцієнти, за допомогою яких можна визначити значення цінної інформації за кожною із шкал. Тобто з використанням розроблених шкал і проводиться аналіз даних, в результаті якого сформовано 17 коефіцієнтів [3], які використовуються в алгоритмах прийняття рішень при виявленні шахрайства. З використанням запропонованих коефіцієнтів і кількісні, і якісні, і масиви кількісних та якісних даних шкалюються до однорідних бінарних значень без втрати інформації, оскільки шкалювання іде по кінцевій цілі.

Особливістю кожного з запропонованих коефіцієнту є те, що кожен з них відповідає одному типу даних та може приймати значення 0 або 1, що дозволяє від різнорідних даних перейти до однорідних даних, тобто і кількісні, і якісні дані шкалюються до однорідних бінарних значень без втрати інформації, оскільки шкалювання іде по кінцевій цілі. Це є важливою особливістю, оскільки після перетворення усіх наявних даних до таких коефіцієнтів, буде можливість подати отримані однорідні дані до будь-якої моделі класифікації. Таким чином, отримаємо можливість використання будь-якої моделі класифікації з використанням усіх вхідних даних, що на даний момент є неможливим. Також, буде можливість відстежити, який з коефіцієнтів найбільше вплинув на віднесення користувача до певного класу, що дасть можливість дати відповідь на питання, чому інтелектуальна система прийняла саме таке рішення.

Для прикладу, розглянемо деякі із запропонованих коефіцієнтів: $k_2$ (куплена / не куплена покупка) – якщо кількість зроблених користувачем покупок $\geq K_{2\_\min}$, то $k_2$ попередньо рівне 1 та означає, що користувач не є шахраєм. Якщо кількість покупок $< K_{2\_\min}$, то $k_2$ попередньо рівне 0.5, що означає, що

користувач є підозрілим. Коефіцієнт $K_{2\_\min}$ у даній роботі визначається на основі експертного опитування. Якщо користувач зробив покупку, яка недоступна для нього, то $k_2$ попередньо рівне 0 та означає, що користувач є шахраєм; $k_1$ (кількість кліків з одного пристрою за хвилину) – важливий коефіцієнт при виявленні шахрайства, проте маючи лише його, не можна однозначно виявити шахрая (аномалію в даних); а також $k_8$ (кількість друзів у соціальній мережі) – якщо користувач прив'язаний до соціальної мережі та в нього є достатня кількість друзів, а саме – $\left(K_{8\_\max} \times K_{8\_opt}; \infty\right)$, то можна зробити перевірку на імена (чи друзі реальні) і фото друзів (аналогічно до визначення коефіцієнта $k_{15}$).

Деякі з коефіцієнтів дозволяють однозначно визначити користувача, що дозволяє сформувати базу даних шахраїв та базу знань з характеристиками шахраїв, а деякі коефіцієнти не дозволяють однозначно визначити клас користувача. Саме такий аналіз даних у процесі тестування розробленої системи, що відповідає основним положенням розглядуваної концепції дозволив здійснити розбиття коефіцієнтів на такі групи з використанням алгоритму 1, що детально представлений у роботі [6]:

– перша група охоплює коефіцієнти, які дозволяють провести попередній аналіз даних, а саме однозначно з множини користувачів визначити шахрайських користувачів, органічних та підозрілих. Дозволяє зробити первинну вибірку на основі Decision Tree. Так наприклад у першу групу входить 13 коефіцієнтів, більшість з яких представлено у роботі [6]. Серед коефіцієнтів першої групи є наприклад $k_2$ (куплена / не куплена покупка), який розглянуто вище. Тобто шкали першої групи дозволяють однозначно визначити клас користувача, що дозволить сформувати базу даних та базу знань з відомими визначеними користувачами в подальшому. Приклад шкал першої групи зображено на рис. 5 [3];

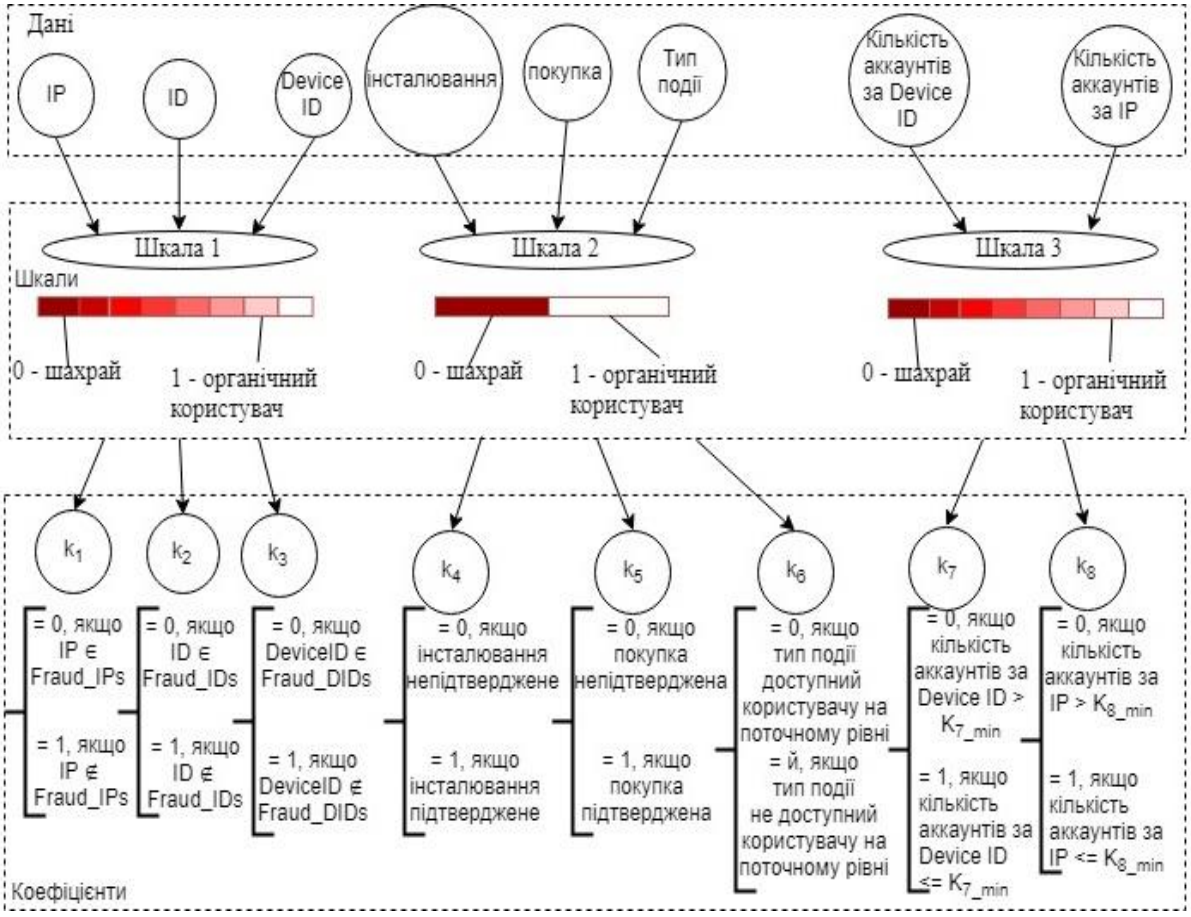



Рисунок 5 – Приклад шкал першої групи

– друга група охоплює коефіцієнти, за якими неможливо зробити первинний аналіз на основі дерева рішень (Decision Tree). Проте, на основі баз даних шахраїв, органічних та підозрілих користувачів та баз знань, сформованих на основі коефіцієнтів з першої групи, можна визначити коефіцієнти схожості усіх користувачів із визначеними користувачами з сформованої бази даних за кожною із характеристик дру-

гої групи. Так, до другої групи відноситься 6 коефіцієнтів, більшість з яких розглянуто у роботі [6]. Серед них можна виділити наприклад $k_1$ (кількість кліків з одного пристрою за хвилину); а також $k_8$ (кількість друзів у соціальній мережі. Слід зазначити, що при визначенні коефіцієнтів другої групи необхідно перевірити усі інші показники. Коефіцієнт $K_{11\_min}$ визначається на основі експертного опитування. Модель шкалювання коефіцієнтів другої групи зображено на рис. 6 [3].

Результатом шкалювання за першою групою представлена таблиця, фрагмент якої представлено у вигляді табл. 1. На основі таких таблиць створюється база даних шахраїв і тому перший етап – це визначення відомих шахраїв по цим таблицям.

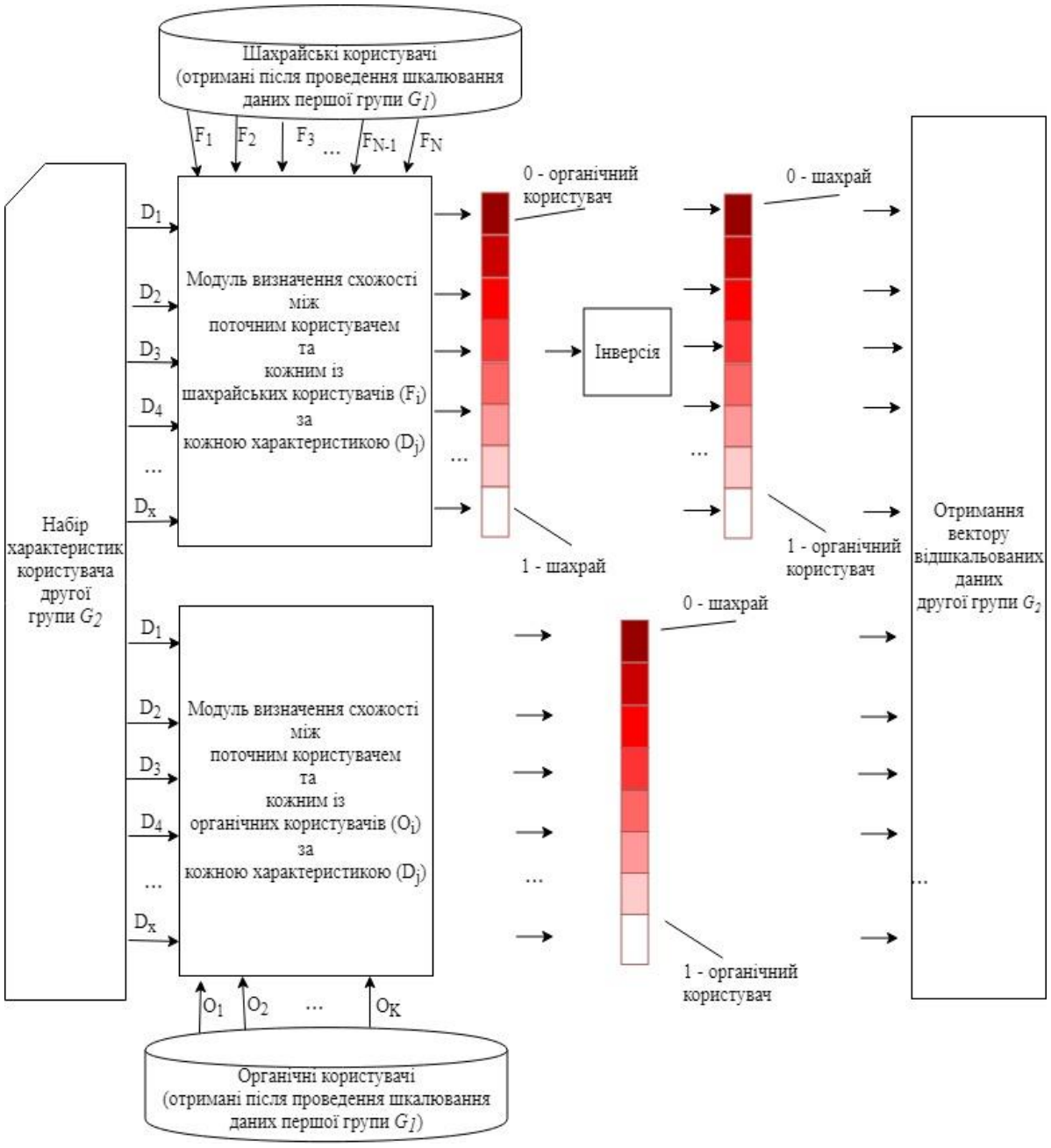



Рисунок 6 – Модель шкалювання даних другої групи

У процесі шкалювання першої групи характеристик і формування коефіцієнтів та формування баз даних і баз знань, можна виділити множини невизначених користувачів. Для цих множин і здійснюється шкалювання за другою групою коефіцієнтів, а саме, здійснюється визначення значень коефіцієнтів другої групи, використовуючи формули визначення коефіцієнтів подібності користувачів.

Таблиця 1 – Приклад принципу роботи запропонованого методу шкалювання

| **Ознака** | **До шкалювання** | **Після шкалювання** | **Коефіцієнт** | **Розпізнавання шахрайства** |
|---|---|---|---|---|
| **IP-адреса** | 127.0.0.1 | 1 | $IP \notin FRAUD\_DeviceID$ | органічний користувач |
| **ID пристрою** | 35 577678 5678735 | 0 | $DeviceID \in FRAUD\_DeviceID$ | шахрай |
| **ID покупки підтверджена чи ні** | 6cf9094a-6fbf-4898-bb3e-0cd895b1cafb | 0 | $PurchaseID.isConfirmed = false$ | шахрай |

Наступним етапом є визначення класів невизначених користувачів. Для цього для кожного з визначених коефіцієнтів другої групи з використанням алгоритму 2, що запропонований у роботі [6], виконуються наступні дії:

1. Визначаються коефіцієнти подібності невизначених користувачів з шахраями із сформованої бази даних, що утворює множину коефіцієнтів, значення яких від 0 до 1.

2. Визначаються коефіцієнти подібності невизначених користувачів з органічними користувачами із сформованої бази даних.

3. Визначаються коефіцієнти подібності невизначених користувачів з підозрілими користувачами із сформованої бази даних.

4. Здійснюється об'єднання отриманих множин значень коефіцієнтів в одну множину однорідних значень.

Зазначимо, що важливим є вибір коефіцієнту подібності, більш детально крок знаходження коефіцієнту подібності описано в роботі [3]. У даній роботі слід відмітити, що значення коефіцієнтів подібності належить проміжку $[0;1]$. Значення 1 означає, що користувачі ідентичні за даною ознакою (feature), 0 у свою чергу означає, що користувачі не мають нічого спільного за даною ознакою (feature). Так, наприклад, розглянувши ознаку по користувачам «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 шахраїв $T_i = \{t \mid t > 0\}$. Маючи дві множини не бінарних, проте однорідних даних $T_u$ та $T_i$, застосуємо відповідний коефіцієнт подібності користувачів. Для множин такого типу, зокрема для множин, що містять інформацію по ознаці «час між подіями», в роботі обрано коефіцієнт Танімото $K_T(T_u, T_i)$ [1,4,12], 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, що використовував, коефіцієнти подібності користувачів визначаються з використанням коефіцієнту косинусної схожості $K_{\cos}(A_1, A_2)$ [1,3,12], 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.

Отримавши множину однорідних значень всіх наявних даних, отриманих з використанням запропонованих шкал, використаємо етап класифікації даних з використанням відомої моделі класифікації (екстремальне градієнтне підсилення XGBoost, випадковий ліс – random forest, глибокі нейронні мережі) з метою виявлення даних, які однозначно визначають шахраїв, і даних, які однозначно визначають органічних користувачів.

**Створення баз знань, що розвиваються**

З метою підвищення ефективності, швидкодії та точності інтелектуальних систем виявлення шахрайства недостатньо лише здійснити класифікацію користувачів на класи «шахрай» чи «органічний», важливим є ще використання баз знань на основі запропонованих вище методів, моделей та алгоритмів.

Крім того, слід зазначити, що в сучасних системах виявлення шахрайства є свої проблеми, однією з яких є відсутність інтелектуального аналізу даних, що не дозволяє таким системам адаптуватися до нових видів шахрайства. Для того, щоб системи мали змогу адаптуватися до нових видів та самонавчатися з використанням алгоритмів прогнозування, автором роботи пропонується система з інтелектуальною складовою – формуванням бази знань, яка дозволить визначити шахраїв, та в яку включатимуться правила аналізу аномалій, при чому так, щоб поява нової аномалії в даних дозволяла створити нове правило. Тобто така база знань може розширюватись через можливість появи нового виду аномалії в даних (шахрайства). Отриманий набір правил, що у подальшому на основі розроблених в роботі алгоритмів дозволить створити узагальнений портрет шахрая, відзначивши навіть нові та невідомі експертам шахрайські властивості. Віднесення підозрілих користувачів до класу шахраїв або органічних відбувається з використанням нечіткої логіки.

Зазначимо, що зв'язок між функціями належності входу $i_j$ з бази знань можна визначати нечіткими логічними рівняннями, сформованими у базі знань, виду:

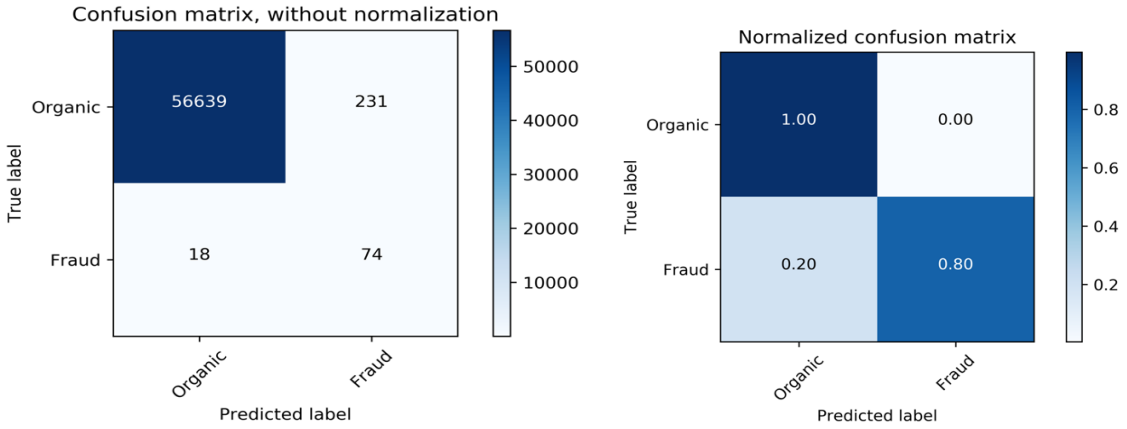
$$\mu^{d_j}(y) = b_{j1}\left[\mu^{j1}(i_1) \wedge \mu^{j1}(i_2) \wedge \ldots \wedge \mu^{j1}(i_n)\right] \vee b_{j2}\left[\mu^{j2}(i_1) \wedge \mu^{j2}(i_2) \wedge \ldots \wedge \mu^{j2}(i_n)\right] \ldots$$
$$b_{jp}\left[\mu^{jp}(i_1) \wedge \mu^{jp}(i_2) \wedge \ldots \mu^{jp}(i_n)\right],$$

$j = \overline{1,m}$, які можна спростити до виразу $\mu^{d_j}(y) = \max_{p=1,k_j}\left\{a_{jp} \min_{i=1,n}\left[\mu^{jp}(i_j)\right]\right\}$, $j = \overline{1,m}$, 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 $b_i^p$, $p = k1$, $i = \overline{1,n}$, $j = \overline{1,m}$. Слід зазначити, що такі бази знань повинні розвиватися в залежності від нових даних та нових шахраїв з метою створення узагальненого портрету шахрая.

**Класифікація користувачів з використанням глибоких мереж**

Наступним етапом в методиці виявлення шахрайства є класифікація отриманої вибірки користувачів на органічних та шахрайських. Але оскільки усі моделі класифікації працюють з однорідними даними, то розробка методу подолання різнорідності даних була необхідною. Тому після етапу подолання різнорідних даних, отримавши відшкальовані та нормалізовані однорідні дані у межах від 0 до 1, може бути використана будь-яка з відомих моделей класифікації. З точки зору авторів для розв'язання такої задачі зручно використовувати моделі класифікації з використанням глибоких нейронних мереж (DNN – deep neural network). Необхідно зазначити, що використання моделей класифікації для поділу користувачів на два визначених класи є відомим та поширеним підходом. На наш погляд для розв'язання такої задачі достатньо використовувати DNN з трьома прихованими шарами.

Маючи множину однорідних значень, здійснимо тренування класифікаційної моделі на вибірці з розробленого мобільного додатку [1, 13-14]. Необхідно зауважити, що для коректного навчання та оцінки системи, а також для уникнення перенавчання системи, необхідно мати три набори даних: тренувальний, перевірочний (валідаційний) та контрольний. Також необхідно відзначити, що насамперед розроблена система повинна бути наділена прогностичною здатністю та добре узагальнюватися на нових для неї даних. Контрольна вибірка містить 56962 записи, 56657 з яких відповідають органічним користувачам, а 305 відповідають користувачам-шахраям. Вибірки помічені, проте мітка використовується лише для подальшої оцінки адекватності моделі. У якості класифікаційної моделі обрано глибокі нейронні мережі із трьома прихованими шарами. На рис. 7 представлено побудовано матрицю невідповідності з використанням мови Python для оцінки адекватності реалізованої класифікаційної моделі. З рисунку 7б видно, що практично 100% органічних користувачів та 80% користувачів-шахраїв визначено правильно.




а б

Рисунок 7 – Матриця невідповідності контрольної вибірки: а – без нормалізації, б – нормалізована

На основі запропонованих в роботі методів, моделей та алгоритмів була розроблена узагальнена структура інтелектуальної системи виявлення шахрайства (рис. 8), що складається з наступних підсистем: підсистема виявлення характеристик користувача; підсистема подолання різнорідності даних; підсистема тренування класифікаційної моделі; підсистема класифікації; підсистема формування бази даних шахраїв; підсистема формування бази знань (для виявлення шахраїв); підсистема інтелектуального аналізу даних та формування шаблонів користувачів; підсистема прогнозування узагальненого шаблону шахрая. Слід зазначити, що це є базовий набір підсистем запропонованої узагальненої структури інтелектуальної системи виявлення шахрайства та можливе розширення для прогнозування, передбачення, коли і звідки прийде наступний шахрай.

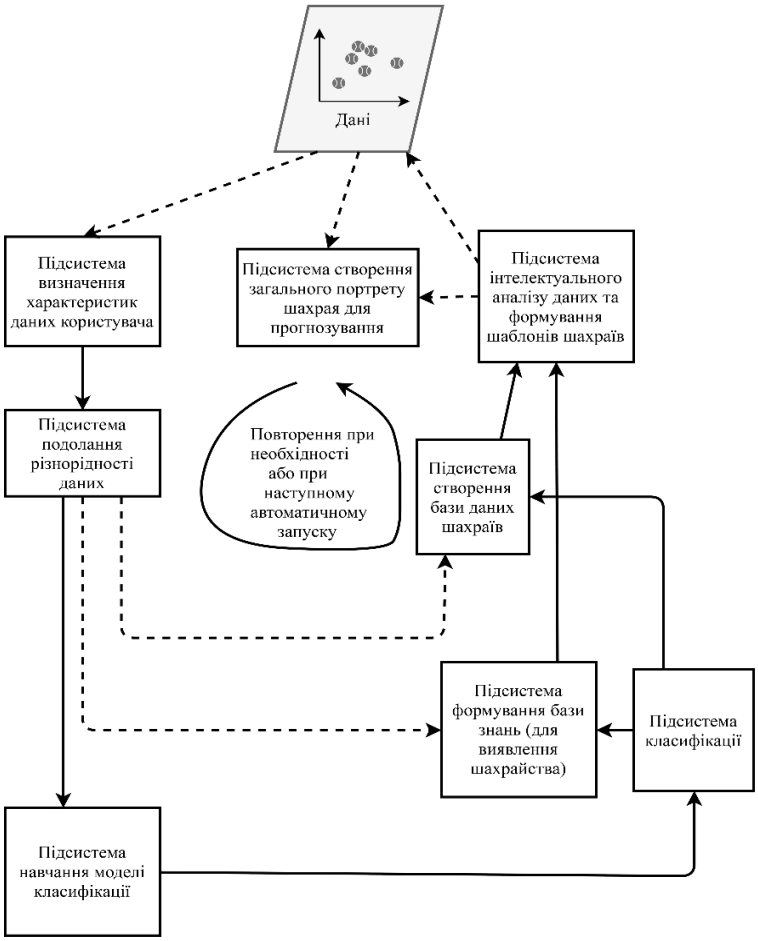



Рисунок 8 – Структура інтелектуальної системи виявлення шахрайства при інсталюванні мобільних додатків

**Висновки**

Отже, у роботі запропоновано інформаційну технологію побудови інтелектуальних систем виявлення шахрайства при інсталюванні мобільних додатків, яка включає:

– запропонований метод аналізу даних, який на відміну від існуючих дозволяє на базі запропонованих шкал подолати різнорідність даних та здійснити інтелектуальну обробку наявних даних з використанням запропонованих алгоритмів 1 та 2 (шкалювання першої та другої групи), що дозволило суттєво підвищити точність отриманих результатів прийняття рішень по виявленню шахраїв;

– метод класифікації користувачів системи, який на відміну від існуючих використовує моделі глибоких нейронних мереж та запропонований метод подолання різнорідності, що дозволило створити бази знань, що розвиваються, з набором нечітких правил та дозволило здійснювати створення узагальненого портрету шахрая в залежності від нових даних та нових шахраїв;

– проведено оцінку адекватності запропонованої класифікаційної моделі на вибірці з розробленого мобільного додатку [1, 13-14], що показала, що практично 100% органічних користувачів та 80% користувачів-шахраїв визначено правильно.

Усі вище розглянуті метод, модель та алгоритми, розроблені для вирішення виділених на початку статті задач, представляються собою технологію.

**Список літератури**


[1] А. А. Яровий, О. Н. Романюк, І. Р. Арсенюк, 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.

[2] Т. Д. Польгуль, А. А. Яровий «Визначення шахрайських операцій при встановленні мобільних додатків з використанням інтелектуального аналізу даних», на *Сучасні тенденції розвитку системного програмування. Тези доповідей,* Київ, 2016, с. 55–56.

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

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

[5] FraudScore: FraudScore fights ad fraud using Machine Learning [Online]. Available: https://fraudscore.mobi/

[6] T. Polhul, 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. doi: 10.15587/1729-4061.2019.155060

[7] V. Chandola, A. Banerjee, V. Kumar, "Anomaly Detection: A Survey", *ACM Computing Surveys (CSUR)*, Volume 41, Issue 3, Article No. 15, New York, NY, USA, July 2009.

[8] I. S. Ivaskiv "Machine-learning methods in tasks of detection the atypical behavior of complex system", *Master Thesis*. Ternopil National Economy University, Ternopil, 2017. (Ukr.).

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

[10] V. Barnett, T. Lewis, *Outliers in Statistical Data*, Wiley, 1994.

[11] A. Géron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, O'Reilly Media, 2017, 574 p.

[12] А. Г. Кюльян, Т. Д. Польгуль, М. Б. Хазін, «Математична модель рекомендаційного сервісу на основі методу колаборативної фільтрації», *Комп'ютерні технології та Інтернет в інформаційному суспільстві,* с. 226–227, 2012.

[13] А. А. Яровий, 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.

[14] А. А. Яровий, 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.


**References**


[1] A. A. Yarovyi, O. N. Romaniuk, I. R. Arseniuk, T. D. Polhul «Vyiavlennia shakhraistva pry instaliuvanni prohramnykh dodatkiv z vykorystanniam intelektualnoho analizu danykh», Naukovi pratsi Donetskoho natsionalnoho tekhnichnoho universytetu. Seriia: «Informatyka, kibernetyka ta obchysliuvalna tekhnika», № 2 (25), с. 126–131, 2017.

[2] T. D. Polhul, A. A. Yarovyi «Vyznachennia shakhraiskykh operatsii pry vstanovlenni mobilnykh dodatkiv z vykorystanniam intelektualnoho analizu danykh», na Suchasni tendentsii rozvytku systemnoho prohramuvannia. Tezy dopovidei, Kyiv, 2016, s. 55–56.
[3] T. D. Polhul, A. A. Yarovyi «Metod podolannia riznoridnosti danykh dlia vyiavlennia shakhraistva pry instaliuvanni mobilnykh dodatkiv», Visnyk SNU im. V. Dalia – Sievierodonetsk: SNU im. V. Dalia. 2018, № 7 ( 248 ), c.60-69, 2018.
[4] T. Polhul, "Development of an intelligent system for detecting mobile app install fraud", in Proceedings of the IRES 156th International Conference, Bangkok, Thailand, 21st-22nd March 2019, pp. 25-29.
[5] FraudScore: FraudScore fights ad fraud using Machine Learning [Online]. Available: https://fraudscore.mobi/
[6] T. Polhul, 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. doi: 10.15587/1729-4061.2019.155060
[7] V. Chandola, A. Banerjee, V. Kumar, "Anomaly Detection: A Survey", ACM Computing Surveys (CSUR), Volume 41, Issue 3, Article No. 15, New York, NY, USA, July 2009.
[8] I. S. Ivaskiv "Machine-learning methods in tasks of detection the atypical behavior of complex system", Master Thesis. Ternopil National Economy University, Ternopil, 2017. (Ukr.).
[9] D. Hawkins, Identification of Outliers. Chapman and Hall, 1980.
[10] V. Barnett, T. Lewis, Outliers in Statistical Data, Wiley, 1994.
[11] A. Géron, Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, OReilly Media, 2017, 574 p.
[12] A. H. Kiulian, T. D. Polhul, M. B. Khazin, «Matematychna model rekomendatsiinoho servisu na osnovi metodu kolaboratyvnoi filtratsii», Kompiuterni tekhnolohii ta Internet v informatsiinomu suspilstvi, s. 226–227, 2012.
[13] A. A. Yarovyi, T. D. Polhul, Kompiuterna prohrama «Prohramnyi modul zboru danykh informatsiinoi tekhnolohii» vyiavlennia shakhraistva pry instaliuvanni prohramnykh dodatkiv. Cvidotstvo pro reiestratsiiu avtorskoho prava na tvir № 76348. K.: Ministerstvo ekonomichnoho rozvytku i torhivli Ukrainy, 2018.
[14] A. A. Yarovyi, T. D. Polhul, Kompiuterna prohrama «Prohramnyi modul vyznachennia skhozhosti korystuvachiv informatsiinoi tekhnolohii vyiavlennia shakhraistva pry instaliuvanni prohramnykh dodatkiv». Cvidotstvo pro reiestratsiiu avtorskoho prava na tvir № 76347. K.: Ministerstvo ekonomichnoho rozvytku i torhivli Ukrainy, 2018.

Стаття надійшла: 05.04.19.

**Відомості про авторів**

**Польгуль Тетяна Дмитрівна** – аспірант кафедри Комп'ютерних наук ВНТУ.

T. D. Polhul

# INFORMATION TECHNOLOGY FOR THE CONSTRUCTION OF INTELLIGENT SYSTEMS FOR DETECTING FRAUD DURING MOBILE APPLICATIONS INSTALLATION

Vinnytsia national technical University, Vinnytsia

Т. Д. Польгуль

# ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ ПОСТРОЕНИЯ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ ОПРЕДЕЛЕНИЯ МОШЕННИЧЕСТВА ПРИ ИНСТАЛЛИРОВАНИИ МОБИЛЬНЫХ ПРИЛОЖЕНИЙ

Винницкий национальный технический университет, Винница