УДК 004.9:004.62

Д. Ю. Тавров, О. Р. Чертов
Національний технічний університет України
“Київський політехнічний інститут”,
кафедра прикладної математики

ЗАБЕЗПЕЧЕННЯ ГРУПОВОЇ АНОНІМНОСТІ ЯК ЗАДАЧА ПОШУКУ ПОТОКУ В МЕРЕЖІ МІНІМАЛЬНОЇ ВАРТОСТІ

Показано, що задачу забезпечення групової анонімності даних можна розглядати як узагальнену задачу пошуку потоку мінімальної вартості в мережі, на архітектуру якої накладено нечіткі обмеження. Для розв’язання цієї задачі запропоновано нову інформаційну технологію. Результати застосування технології проілюстровано за допомогою прикладу на основі реальних даних.

Ключові слова: групова анонімність даних, мікрофайл, інформаційна технологія.

In the paper, it is shown that the task of providing group anonymity can be treated as a generalized minimum cost flow problem, where fuzzy restrictions are imposed on the network architecture. To solve this task, a novel information technology is proposed. Results of applying this technology are illustrated with a real data based example.

Key words: data group anonymity, microfile, information technology.

Вступ. Загальна постановка проблеми

Оприлюднення первинних даних переписів, статистичних спостережень, соціологічних опитувань тощо пов’язане з ризиком розкриття чутливої інформації. Відповідні загрози загальновідомі в галузі збереження приватності публічних даних (privacy-preserving data publishing) [1], у межах якої забезпечують анонімність індивідуальних респондентів, тобто їхню властивість бути нерозрізненними в наборі даних. Останнім часом набув поширення підхід до забезпечення групової анонімності даних, пов’язаний із захистом інформації про групи респондентів [2], що передбачає захист чутливих особливостей розподілу даних про певну групу.

Найпростіший спосіб забезпечення анонімності полягає у вилученні з набору даних ідентифікаторів, тобто атрибутів, що дають змогу однозначно встановити особу респондента, або сутнісних атрибутів, тобто атрибутів, що однозначно визначають групу респондентів. Проте такий підхід не гарантує досконалого захисту даних. Зокрема, у галузі забезпечення індивідуальної анонімності вилучення ідентифікаторів недостатнє, оскільки особу респондента можна визначити, використовуючи так звані квазіідентифікатори [3] – атрибути, комбінації значень яких дають змогу розкрити особу респондента з достатнім ступенем упевненості. Для забезпечення групової анонімності вилучення сутнісних атрибутів також не є достатнім, оскільки, як було показано в [4], існує можливість побудови нечіткої моделі групи у вигляді системи нечіткого виведення, за допомогою якої одержують наближення розподілу даних про групу.

Підходи до забезпечення групової анонімності можна розділити на дві групи [5]: двоетапні та одноетапні. Усі двоетапні підходи передбачають виконання двох кроків:

1. Одержати модифікований розподіл, який маскує чутливі до розкриття особливості групи.
2. Визначити процедуру модифікації даних із метою увідповіднення їх до модифікованого розподілу.

На другому етапі, як правило, у дані бажано вносити спотворення якомога меншого обсягу. У цій роботі показано, що модифікацію даних зі спотвореннями мінімального обсягу можна розглядати як задачу пошуку потоку в мережі мінімальної вартості, відому в теорії оптимізації потоків у мережі [6, с. 296]. Між архітектурою мережі та модифікованим розподілом існує взаємно

однозначна відповідність. Для розв'язання такої задачі в літературі описано низку алгоритмів поліноміальної та псевдополіноміальної складності.

Якщо на модифікований розподіл даних про групу не накладено додаткових обмежень, можна одержати розподіли, еквівалентні з погляду маскування властивостей даних. Наприклад, у випадку маскування територіального розподілу військовослужбовців, викиди в якому можуть указувати на місця розташування прихованих військових об'єктів, групову анонімність можна забезпечити модифікацією даних у такий спосіб, що структура викидів розподілу буде іншою. Наприклад, можна створити нові викиди у певних регіонах, знизивши значення початкових викидів до такого рівня, що вони будуть нерозрізненні в модифікованому розподілі. При цьому вибір регіонів для створення нових викидів може бути довільним, а два модифіковані розподіли можуть відрізнятися абсолютними значеннями, навіть якщо вибір регіонів для них був однаковий. Очевидно, що різні модифіковані розподіли можуть бути еквівалентними з погляду забезпечення групової анонімності на першому етапі, але відповідати різним архітектурам мережі, тобто на другому етапі різні модифіковані розподіли іноді призводять до одержання спотворень різного обсягу.

На відміну від двоетапного, одноетапний підхід до забезпечення групової анонімності передбачає одержання модифікованого розподілу, який, з одного боку, задовольняє вимогу маскування чутливих властивостей даних, а з іншого, призводить до внесення в дані спотворень мінімального обсягу. Модифікацію даних відповідно до одноетапного підходу можна формалізувати у вигляді задачі пошуку потоку в мережі мінімальної вартості з додатковими умовами, накладеними на архітектуру мережі, що враховують вимогу щодо якості модифікованого розподілу з погляду маскування чутливих властивостей даних. З урахуванням суб'єктивної та неточної природи таких обмежень, їх формалізують у вигляді нечітких обмежень (fuzzy restrictions) [7], накладених на шуканий розподіл.

Узагальнена задача пошуку потоку мінімальної вартості складніша за класичну, оскільки:

– у більшості практичних випадків нечіткі обмеження мають нелінійну природу;
– анонімізовані дані, як правило, належать до даних великого обсягу.

За час розвитку галузі анонімізації даних розроблено низку програмних засобів для забезпечення як індивідуальної, так і групової анонімності даних. Історично першими системами, які реалізовували окремі алгоритми анонімізації даних, були системи Scrub [8] та Datafly [9], розроблені Л. Суїні. Сьогодні найпотужнішим є пакет прикладних програм μ-ARGUS [10], розроблений відповідно до проектів SDC (Statistical Disclosure Control) Четвертої рамкової програми Європейського Союзу та CASC (Computational Aspects of Statistical Confidentiality) П'ятої рамкової програми Європейського Союзу. Тепер проект розвивається за підтримки Євростату. Пакет μ-ARGUS дає можливість виконувати індивідуальну анонімізацію мікрофайлів та поширюється у вільному доступі, але не підтримує методів забезпечення групової анонімності.

У роботі [11] реалізовано окремі методи забезпечення групової анонімності, але в цій реалізації не враховується вплив значень атрибутів, які не є сутнісними, на можливість порушення анонімності групи, а на етапі оцінювання якості розв'язку ЗЗГА обов'язковою є участь експерта.

Метою цієї роботи є зведення задачі забезпечення групової анонімності до задачі пошуку потоку в мережі мінімальної вартості та розроблення на основі цього нової інформаційної технології забезпечення групової анонімності даних, яка враховує комбінації значень атрибутів мікрофайла, що не є сутнісними, і не потребує участі експерта в оцінюванні якості одержуваних розв'язків. Ця інформаційна технологія повинна задовольняти такі вимоги:

1. Мати зручний у користуванні графічний інтерфейс із можливістю подання даних у графічній та табличній формах.
2. Бути портовною, тобто не залежати від платформи та операційного середовища.
3. Не вимагати від користувача професійної підготовки в галузі інформаційних технологій.
4. Процеси обробки даних повинні бути гнучкими, із можливістю налагодження параметрів.

Основні поняття групової анонімності

Нехай маємо (деперсоніфікований) *мікрофайл* $\mathbf{M}$ із даними, анонімність яких потрібно забезпечити. Мікрофайл складається із *записів* $\mathbf{r}^{(i)}$, $i=\overline{1,\rho}$, які містять значення z_{ij} *атрибутів* w_j, $j=\overline{1,\eta}$. Множину значень атрибута w_j позначатимемо через $\mathbf{w}_j$. *Групу* записів у мікрофайлі можна визначити за допомогою двох множин значень атрибутів:

– множини $\mathbf{V}=\{V_1,V_2,\ldots,V_{l_v}\}$ комбінацій значень *сутнісних* атрибутів w_{v_j}, $j=\overline{1,l}$, які відображають характеристики записів, за допомогою яких можна однозначно класифікувати запис як належний групі. Кожна *сутнісна комбінація значень* V_i, $i=\overline{1,l_v}$, – елемент декартового добутку $\mathbf{w}_{v_1}\times\mathbf{w}_{v_2}\times\ldots\times\mathbf{w}_{v_l}$. Записи $\mathbf{M}$, значення атрибутів яких належать $\mathbf{V}$, називатимемо *сутнісними*;

– множини $\mathbf{P}=\{P_1,P_2,\ldots,P_{l_p}\}$ значень *параметризувального* атрибута w_p, $p\neq v_j\ \forall j$, яка задає значення, за якими потрібно виконати розподіл даних про групу. *Параметризувальні значення* $P_i\in\mathbf{w}_p$, $i=\overline{1,l_p}$ дають змогу розділити мікрофайл $\mathbf{M}$ на *параметричні підмікрофайли* $\mathbf{M}_1,\ldots,\mathbf{M}_{l_p}$, кожний із яких містить ρ_k записів, $k=\overline{1,l_p}$, $\sum_k\rho_k=\rho$.

Усі інші атрибути w_{b_j}, $j=\overline{1,t}$ називатимемо *базовими атрибутами*.

Кожну групу позначатимемо через $G(\mathbf{V},\mathbf{P})$. Розподіл даних про неї, який потрібно анонімізувати, називатимемо *цільовим поданням* (ЦП) і позначатимемо через $\Omega(\mathbf{M},G)$. У цій роботі розглядатимемо подання у вигляді *цільового сигналу* $\boldsymbol{\theta}=(\theta_1,\theta_2,...,\theta_{l_p})$ яке найчастіше використовують на практиці. Цільовий сигнал може бути *кількісним* $\mathbf{q}=(q_1,q_2,\ldots,q_{l_p})$, де q_k, $k=\overline{1,l_p}$, – загальна кількість сутнісних записів у $\mathbf{M}_k$, або *концентраційним* $\mathbf{c}=(c_1,c_2,\ldots,c_{l_p})$, елементи якого одержують, поділивши q_k на загальну кількість записів у відповідному підмікрофайлі. У роботі розглядатимемо тільки кількісний сигнал, оскільки всі розглядувані методи та моделі можна очевидним способом узагальнити на випадок концентраційного сигналу.

Під чутливими особливостями кількісного сигналу, які потрібно захистити, розумітимемо його *викиди*, які можна визначити за допомогою процедури на основі *модифікованого методу* τ *Томпсона* (ММТТ), який у стандарті ASME PTC 19 рекомендується [12, с. 79] застосовувати на практиці для визначення викидів. Ця процедура передбачає виконання таких кроків:

а) впорядкувати значення $\mathbf{q}$ за зростанням;

б) обчислити медіану та псевдосередньоквадратичне відхилення сигналу:

$$M_{\mathbf{q}}=\begin{cases}q_{(m_{\mathbf{q}}+1)/2}, & m_{\mathbf{q}}-\text{непарне}\\ \dfrac{q_{m_{\mathbf{q}}/2}+q_{m_{\mathbf{q}}/2+1}}{2}, & m_{\mathbf{q}}-\text{парне}\end{cases},\quad s_{ps\mathbf{q}}=\frac{q_{0,75}-q_{0,25}}{1,349}, \tag{1}$$

де $m_{\mathbf{q}}$ – кількість елементів у $\mathbf{q}$; $q_{0,75}$ ($q_{0,25}$) – верхня (нижня) квартиль $\mathbf{q}$. Якщо $m_{\mathbf{q}}$ парне, то верхня (нижня) квартиль – медіана найбільших (найменших) $m_{\mathbf{q}}/2$ значень $\mathbf{q}$. Якщо $m_{\mathbf{q}}$ непарне, верхня (нижня) квартиль – медіана найбільших (найменших) $(m_{\mathbf{q}}+1)/2$ значень $\mathbf{q}$;

в) обчислити для кожного i-го значення сигналу $\mathbf{q}$ його абсолютне відхилення від $M_{\mathbf{q}}$:

$$d_i=\left|q_i-M_{\mathbf{q}}\right|; \tag{2}$$

г) обчислити значення τ за формулою

$$\tau = \frac{t_{\alpha/2} \cdot \left(m_{\mathbf{q}} - 1\right)}{\sqrt{m_{\mathbf{q}}}\sqrt{m_{\mathbf{q}} - 2 + t_{\alpha/2}^2}}, \tag{3}$$

де $t_{\alpha/2}$ – критичне значення t розподілу Стьюдента [13], обчислене для рівня значущості α та кількості ступенів свободи $m_{\mathbf{q}} - 2$;

д) якщо $\exists i \; d_i > \tau s_{ps\mathbf{q}}$, то i-те значення є викидом, і його потрібно вилучити з сигналу та повернутися на крок *б*. Якщо $\forall i \; d_i \le \tau s_{ps\mathbf{q}}$, алгоритм припиняє свою роботу.

Позначмо через $OUT(\mathbf{q})$ множину індексів елементів $\mathbf{q}$, які відповідають викидам, визначеним за допомогою ММТТ. Позначмо через $OUT_e(\mathbf{q}) \subseteq OUT(\mathbf{q})$ множину індексів елементів $\mathbf{q}$, утворену вилученням з $OUT(\mathbf{q})$ індексів, які, на думку експерта, не відповідають викидам, які потрібно маскувати, забезпечуючи анонімність.

Задача забезпечення групової анонімності (ЗЗГА) полягає у визначенні такої модифікації $\mathbf{M}$ із метою одержання захищеного мікрофайла $\mathbf{M}^*$, яка дає змогу маскувати викиди кількісного сигналу і при цьому гарантувати мінімальний обсяг спотворень, який вноситься в дані.

Найпростішим розв'язком ЗЗГА є вилучення з мікрофайла сутнісного атрибута. Проте навіть якщо його буде вилучено, залишається можливість побудови *нечіткої моделі групи* у вигляді системи нечіткого виведення (СНВ). Аналізуючи таку модель та відповідний їй ЦП, групову анонімність можна порушити. Вхідними змінними нечіткої моделі повинні бути лінгвістичні змінні [14], які відповідають базовим атрибутам. Тоді можна визначити *ступінь належності* $\mu_G\left(\mathbf{r}^{(i)}\right)$ кожного запису $\mathbf{r}^{(i)} \in \mathbf{M}$ групі G. Розрізняють два основні типи нечітких моделей:

– нечітка модель на основі сторонніх даних [15], яку можна збудувати у разі наявності доступу до допоміжного мікрофайла $\tilde{\mathbf{M}}$, близького до мікрофайла $\mathbf{M}$, на думку експерта, як за структурою, так і за вмістом. У $\tilde{\mathbf{M}}$, на відміну від $\mathbf{M}$, містяться атрибути зі значеннями, інтерпретація яких ідентична до інтерпретації сутнісних комбінацій значень $\mathbf{M}$. ЦП мікрофайла відносно такої нечіткої моделі є *допоміжне цільове подання* (ДЦП) $\tilde{q}_j = \sum_{\mathbf{r} \in \mathbf{M}_j | \mu_G(\mathbf{r}) \ge \alpha} \mu_G(\mathbf{r})$. Базу нечітких правил моделі будують за допомогою еволюційного алгоритму, описаного в [15];

– нечітка модель на основі експертних знань [4], яку можна збудувати у випадку відсутності доступу до допоміжного мікрофайла. ЦП мікрофайла відносно такої моделі є *узагальнене цільове подання* (УЦП) у вигляді кількісної поверхні $\mathbf{Q}$ з елементами $Q_{sk} = \left| \left\{ \mathbf{r}^{(i)} \middle| z_{iw_p} = P_k, \mu_G\left(\mathbf{r}^{(i)}\right) \in \Delta_s \right\} \right|$, де Δ_s, $s = \overline{1, l_{int}}$, – інтервали значень ступенів належності респондентів групі.

Щоб модифікувати кількісний сигнал (або, у випадку можливості побудови нечітких моделей, ДЦП чи УЦП), потрібно змінити параметризувальні значення певних записів мікрофайла. Щоб забезпечити схоронність загальної кількості записів у кожному підмікрофайлі, записи потрібно модифікувати парами, тобто фактично обмінювати записи між підмікрофайлами. При цьому в дані повинно бути внесено спотворення якомога меншого обсягу. Це означає, що обмінювати потрібно записи, близькі в деякому розумінні. У цій роботі близькість двох записів $\mathbf{r}$ та $\mathbf{r}^*$ визначатимемо за допомогою *визначальної метрики* [16], яку описують у термінах *визначальних атрибутів*:

$$\mathrm{InfM}(\mathbf{r}, \mathbf{r}^*) = \sum_{p=1}^{n_{ord}} \omega_p \left(\frac{r_{I_p} - r^*_{I_p}}{r_{I_p} + r^*_{I_p}} \right)^2 + \sum_{k=1}^{n_{cat}} \gamma_k \chi^2\left(r_{J_k}, r^*_{J_k}\right), \tag{4}$$

де I_p – p-й порядковий визначальний атрибут (розподіл значень якого важливий для подальших досліджень із використанням даних мікрофайла), загальна кількість таких атрибутів – n_{ord};

J_k – k-й категорійний визначальний атрибут, загальна кількість таких атрибутів – n_{cat};

$\chi(v_1, v_2)$ – оператор, який дорівнює χ_1, якщо значення v_1 та v_2 належать одній категорії, та χ_2 – у протилежному випадку;

ω_p й γ_k – невід'ємні ваги (що більша вага, то важливіший атрибут для досліджень).

Задача забезпечення групової анонімності як задача пошуку потоку в мережі мінімальної вартості

Розгляньмо спочатку постановку задачі пошуку потоку в мережі мінімальної вартості для двоетапного підходу до розв'язання ЗЗГА. Позначмо через δ_i *валентність* параметричного підмікрофайла $\mathbf{M}_i$, $i = \overline{1, l_p}$. Валентності показують, скільки записів потрібно додати в підмікрофайл (якщо $\delta_i > 0$) або вилучити (якщо $\delta_i < 0$). Очевидно, що $\delta_i = q_i - q_i^*$, $i = \overline{1, l_p}$.

Позначмо через $G = (N, A)$ орієнтовану мережу, визначену множиною N із n вузлів та множиною A з m орієнтованих дуг. Кожна дуга $(i, j) \in A$ характеризується вартістю c_{ij} та пропускною здатністю u_{ij}. Кожний вузол $i \in N$ асоційовано з деяким числом $b(i)$, яке можна інтерпретувати як його пропозицію, якщо $b(i) > 0$, або попит, якщо $b(i) < 0$. Тоді задачу пошуку потоку мінімальної вартості можна сформулювати так:

$$\begin{gathered} \min z(x) = \sum_{(i,j)\in A} c_{ij} x_{ij}, \\ \sum_{j|(i,j)\in A} x_{ij} - \sum_{j|(j,i)\in A} x_{ji} = b(i) \quad \forall i \in N, \\ 0 \le x_{ij} \le u_{ij} \quad \forall i(i,j) \in A, \\ \sum_{i=1}^{n} b(i) = 0. \end{gathered}, \tag{5}$$

У контексті ЗЗГА архітектуру мережі можна визначити так:

– множина N складається з чотирьох підмножин N_1, N_2, N_3 та N_4, що не перетинаються;

– вузли $N_1^k \in N_1$ та $N_4^k \in N_4$ відповідають параметричним підмікрофайлам $\mathbf{M}_k$, $k = \overline{1, l_p}$, тобто $|N_1| = |N_4| = l_p$;

– множину N_2 можна розділити на l_p підмножин $N_2^{(1)}, \ldots, N_2^{(l_p)}$, що не перетинаються. Вузли $N_2^{(k,i)} \in N_2^{(k)}$ відповідають сутнісним записам підмікрофайла $\mathbf{M}_k$, $k = \overline{1, l_p}$, $i = \overline{1, q_k}$, тобто $\left|N_2^{(k)}\right| = q_k$, $k = \overline{1, l_p}$;

– множину N_3 можна аналогічно розділити на l_p підмножин $N_3^{(1)}, \ldots, N_3^{(l_p)}$, що не перетинаються. Вузли $N_3^{(k,i)} \in N_3^{(k)}$ відповідають несутнісним записам підмікрофайла $\mathbf{M}_k$, $k = \overline{1, l_p}$, $i = \overline{1, \rho_k - q_k}$, тобто $\left|N_3^{(k)}\right| = \rho_k - q_k$, $k = \overline{1, l_p}$;

– множина A складається з трьох підмножин $A_{N_1N_2}$, $A_{N_2N_3}$ та $A_{N_3N_4}$, що не перетинаються;

– дуги в $A_{N_1N_2}$ з'єднують вузли $N_1^k \in N_1$ з вузлами $N_2^{(k,i)} \in N_2^{(k)}$, $k = \overline{1, l_p}$, $i = \overline{1, q_k}$. Жодних інших дуг у множині $A_{N_1N_2}$ немає;

– дуги в $A_{N_3N_4}$ з'єднують вузли $N_3^{(k,i)} \in N_3^{(k)}$ з вузлами $N_4^k \in N_4$, $k = \overline{1, l_p}$, $i = \overline{1, \rho_k - q_k}$. Жодних інших дуг у множині $A_{N_3N_4}$ немає;

– дуги в $A_{N_2N_3}$ з'єднують кожний вузол у $N_2^{(k,i)} \in N_2^{(k)}$, $k=\overline{1,l_p}$, $i=\overline{1,q_k}$, із кожним вузлом $N_3^{(l,j)} \in N_3^{(l)}$, $l=\overline{1,l_p}$, $k \neq l$, $j=\overline{1,\rho_l - q_l}$;

– пропозиції вузлів $N_1^k \in N_1$, $k=\overline{1,l_p}$, дорівнюють $b\left(N_1^k\right)=\begin{cases}\delta_k, & \delta_i > 0\\ 0, & \delta_i \le 0\end{cases}$, де δ_k – валентність підмікрофайла $\mathbf{M}_k$;

– попити вузлів $N_4^k \in N_4$, $k=\overline{1,l_p}$, дорівнюють $b\left(N_4^k\right)=\begin{cases}\delta_k, & \delta_i < 0\\ 0, & \delta_i \ge 0\end{cases}$;

– пропозиції/попити вузлів у N_2 та N_3 дорівнюють 0;

– пропускні здатності всіх дуг у A дорівнюють 1;

– вартості дуг у $A_{N_1N_2}$ та $A_{N_3N_4}$ дорівнюють 0;

– вартість кожної дуги $\left(N_2^{(k,i)}, N_3^{(l,j)}\right) \in A_{N_2N_3}$, $k=\overline{1,l_p}$, $l=\overline{1,l_p}$, $k \neq l$, $i=\overline{1,q_k}$, $j=\overline{1,\rho_l - q_l}$, дорівнює значенню метрики (4), обчисленої для відповідної пари записів мікрофайла.

Мережу, описану вище, подано на рис. 1.

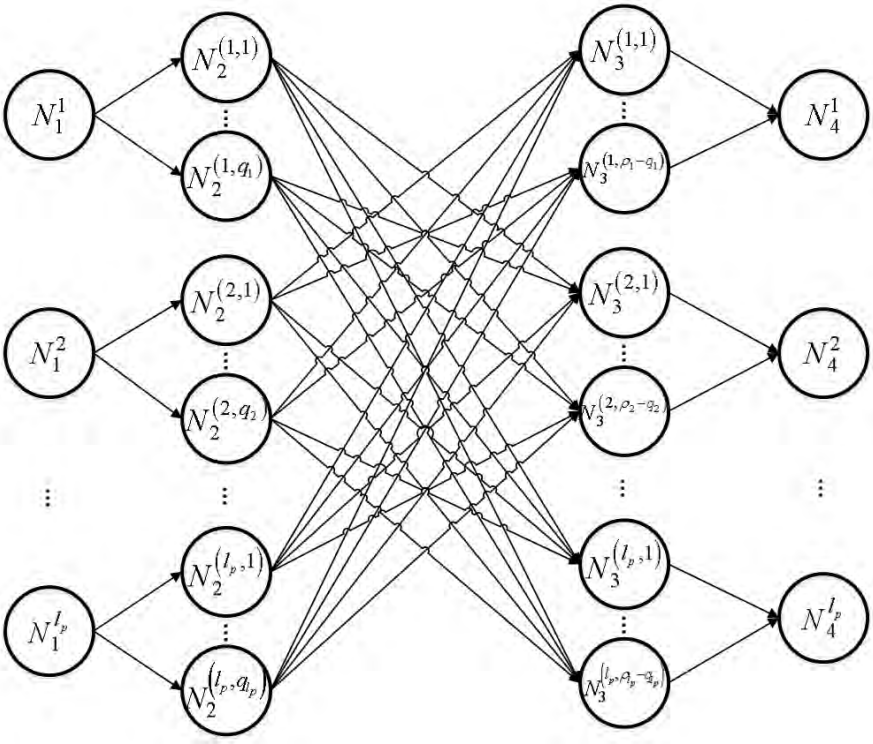

Рис. 1. Архітектура мережі для ЗЗГА

За одноетапного підходу до розв'язання ЗЗГА вигляд модифікованого кількісного сигналу $\mathbf{q}^*$ невідомий, і на його значення можна тільки накласти нечіткі обмеження [7] такого вигляду:

$$\begin{gathered} R\left(X_{i_1}\right): X_{i_1} \text{ is } A_{i_1}, \\ R\left(X_{i_2}\right): X_{i_2} \text{ is } A_{i_2}, \\ \cdots \qquad , \\ R\left(X_{i_k}\right): X_{i_k} \text{ is } A_{i_k}, \\ R\left(X_{\bar{J}} \mid q_J^*\right): \sum_{i \in \bar{J}} q_i^* = \rho_v - \sum_{j \in J} q_j^*. \end{gathered} \tag{6}$$

де $\mathbf{q}^*=\left(q_1^*,\ldots,q_{l_p}^*\right)$ можна інтерпретувати як значення l_p-арної змінної $X=\left(X_1,\ldots,X_{l_p}\right)$; $J=\left(i_1,\ldots,i_k\right)$ – послідовність індексів, яка є власною підпослідовністю $I=\left(1,\ldots,l_p\right)$; $\overline{J}$ – доповнення J до I; X_J – k-арна змінна $X_J=\left(X_{i_1},\ldots,X_{i_k}\right)$.

ЗЗГА можна розглядати як *узагальнену* задачу пошуку потоку в мережі мінімальної вартості:

$$\begin{gathered}\min z(x)=\sum_{(i,j)\in A}c_{ij}x_{ij},\\ \sum_{j|(i,j)\in A}x_{ij}-\sum_{j|(j,i)\in A}x_{ji}=b(i)\quad\forall i\in N,\\ 0\le x_{ij}\le u_{ij}\quad\forall i(i,j)\in A,\\ \sum_{i=1}^{n}b(i)=0,\qquad,\\ \mu\left(q_1^*,\ldots,q_{l_p}^*\right)\ge\alpha_{comp},\\ \frac{\left|OUT_e(\mathbf{q})\cap OUT\left(\mathbf{q}^*\right)\right|}{\left|OUT_e(\mathbf{q})\right|}\le K_{out}\end{gathered}\qquad(7)$$

де $\mu\left(q_1^*,\ldots,q_{l_p}^*\right)$ – ступінь сумісності модифікованого кількісного сигналу $\mathbf{q}^*$ із нечітким обмеженням $R(X)$ у вигляді (6); α_{comp} – деяка константа, яку називатимемо *порогом сумісності*; ступінь сумісності $\mathbf{q}^*$ з $R(X)$ вважають прийнятним, якщо він вищий за α_{comp}; як правило, значення цієї константи потрібно вибирати більшим за 0,5; K_{out} – деяка константа, яку називатимемо *порогом чутливості*; якщо $K_{out}=0$, то розв'язком ЗЗГА вважатиметься модифікація, застосування якої дає змогу одержати модифікований кількісний сигнал, викиди якого не відповідають викидам початкового сигналу: $OUT_e(\mathbf{q})\cap OUT\left(\mathbf{q}^*\right)=0$.

На практиці через природну неточність статистичних даних знаходити оптимальний розв'язок ЗЗГА в постановці (7) недоцільно. Уведемо поняття *допустимого розв'язку ЗЗГА* як такої впорядкованої послідовності $\mathbf{S}=\left(\left(\mathbf{r}^{(i_1)},\mathbf{r}^{(j_1)}\right),\ldots,\left(\mathbf{r}^{(i_Q)},\mathbf{r}^{(j_Q)}\right)\right)$ пар записів, які потрібно обміняти між різними підмікрофайлами, що задовольняє такі умови:

$$\begin{gathered}\mu\left(q_1^*(\mathbf{S}),\ldots,q_{l_p}^*(\mathbf{S})\right)\ge\alpha_{comp},\\ \frac{\left|OUT_e(\mathbf{q})\cap OUT\left(\mathbf{q}^*(\mathbf{S})\right)\right|}{\left|OUT_e(\mathbf{q})\right|}\le K_{out},,\\ \sum_{k=1}^{Q}\mathrm{InfM}\left(\mathbf{r}^{(i_k)},\mathbf{r}^{(j_k)}\right)\le K_{dist}\cdot C_{\max},\end{gathered}\qquad(8)$$

де $\mathbf{q}^*(\mathbf{S})$ – модифікований сигнал, який можна одержати, виконуючи послідовні обміни записів із послідовності $\mathbf{S}$; K_{dist} – деяка константа з проміжку $[0,1]$, яку називатимемо *порогом спотворень*; $C_{\max}$ – найбільше можливе сумарне значення визначальної метрики (4), яку можна обчислити для розв'язуваної ЗЗГА.

Обсяг спотворень вважають прийнятним, якщо сумарне значення $\sum_{k=1}^{Q}\mathrm{InfM}\left(\mathbf{r}^{(i_k)},\mathbf{r}^{(j_k)}\right)$ для цього розв'язку $\mathbf{S}$ не перевищує $K_{dist}\cdot C_{\max}$.

Узагальнена задача пошуку потоку мінімальної вартості складніша за класичну, оскільки:

– у більшості практичних випадків нечіткі обмеження мають нелінійну природу;
– анонімізовані дані, як правило, належать до даних великого обсягу.

Для розв'язання ЗЗГА в такій постановці може бути використано меметичний алгоритм [7], у якому функція пристосованості явно враховує у вигляді добутку відповідних термів як вимогу до якості розв'язку в розумінні маскування викидів (сумісності з накладеними нечіткими обмеженнями), так і вимогу до якості розв'язку в розумінні мінімізації спотворень мікрофайла.

Опис інформаційної технології

Згідно з ДСТУ 2226-93 [17, с. 6], інформаційна технологія – це технологічний процес, предметом перероблення й результатом якого є інформація, під якою, своєю чергою, розуміють [17, с. 8] відомості про суб'єкти, об'єкти, явища та процеси. Згідно з [18], під інформаційними технологіями розуміють будь-які засоби трансформації даних у корисний для досягнення мети системи управління інформаційний продукт. Схоже визначення інформаційної технології наведено в [19], згідно з яким інформаційна технологія – сукупність засобів і методів збору, оброблення та передавання даних (первинної інформації) для одержання інформації нової якості про стан об'єкта, процесу чи явища (інформаційного продукту).

Інформаційна технологія забезпечення анонімності даних про групи, відносно яких існує загроза її порушення, за допомогою аналізу комбінацій значень базових атрибутів мікрофайла, опис якої наведено в цій роботі, є сукупністю моделей та методів для оброблення деякого мікрофайла $\mathbf{M}$, який містить дані про респондентів, із метою одержання модифікованого мікрофайла $\mathbf{M}^*$, у якому забезпечено анонімність заданих груп респондентів.

Інформаційну технологію прийнято подавати у вигляді ієрархічної структури [19] з різними рівнями, наприклад, етапів, операцій та дій. Складові інформаційної технології забезпечення анонімності даних про групи, відносно яких існує загроза її порушення шляхом аналізу комбінацій значень базових атрибутів мікрофайла, представлено на рис. 2 до рівня операцій включно. Прокоментуємо складові цієї інформаційної технології.

На етапі *побудови моделі групи* користувач повинен визначити групу, анонімність якої потрібно забезпечити, та збудувати відповідне їй ЦП. Для цього потрібно виконати такі операції:

– визначення групи, що передбачає виконання трьох дій: завантаження мікрофайла та файла метаданих, визначення в мікрофайлі сутнісних атрибутів та їхніх значень, визначення в мікрофайлі параметризувального атрибута і його значень та впорядкування цих значень;
– побудова ЦП, що передбачає виконання двох дій: вибір типу ЦП (кількісний або концентраційний сигнал), обчислення значень сигналу;
– визначення викидів у ЦП, що передбачає виконання двох дій: застосування до ЦП ММТТ, вилучення з одержаної множини індексів, що не відповідають викидам.

Рис. 2. Складові інформаційної технології забезпечення групової анонімності даних

На рис. 3 представлено UML-діаграму діяльності (activity diagram) [20] користувача інформаційної технології на етапі побудови моделі групи.

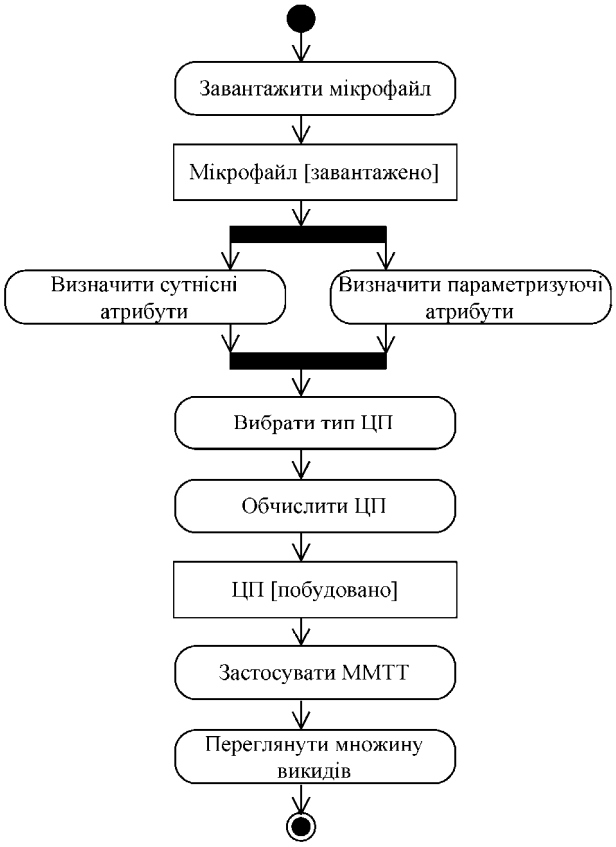

Рис. 3. UML-діаграма активності користувача інформаційної технології на етапі побудови моделі групи

На етапі *побудови нечіткої моделі групи на основі сторонніх даних* користувач повинен збудувати відповідну нечітку модель групи та встановити факт наявності загрози порушення групової анонімності. Для цього потрібно виконати такі операції:

– визначення допоміжного мікрофайла, що передбачає виконання трьох дій: вибір допоміжного мікрофайла (якщо неможливо вибрати, потрібно перейти до виконання наступного етапу), завантаження цього мікрофайла, гармонізація основного та допоміжного мікрофайлів;

– ідентифікація вхідних змінних СНВ, що передбачає виконання трьох дій: визначення проміжку допустимих значень для кожного базового атрибута, вилучення з гармонізованих мікрофайлів записів, значення атрибутів яких лежать поза визначеними проміжками, задання для кожної лінгвістичної змінної, що відповідає базовому гармонізованому атрибуту, нечітких значень;

– побудова бази правил СНВ за допомогою відповідного еволюційного алгоритму [15];

– встановлення факту наявності загрози порушення групової анонімності, що передбачає виконання чотирьох дій: побудова ДЦП, застосування до ДЦП ММТТ, вилучення з одержаної множини індексів, що не відповідають викидам, оцінювання адекватності моделі.

На рис. 4 подано UML-діаграму діяльності користувача на цьому етапі.

– на етапі *побудови нечіткої моделі групи на основі експертних знань* користувач повинен збудувати відповідну нечітку модель групи та встановити факт наявності загрози порушення групової анонімності. Якщо на попередньому етапі встановлено факт наявності загрози порушення анонімності за допомогою нечіткої моделі на основі сторонніх даних, поточний етап потрібно пропустити. Для побудови нечіткої моделі необхідно виконати такі операції:

– визначення параметрів СНВ, що передбачає виконання чотирьох дій: визначення множини атрибутів мікрофайла, які можна використати як вхідні змінні СНВ, задання значень вхідних та вихідної змінної СНВ, побудова множини нечітких правил, вибір методів нечіткого об'єднання, перетину, імплікації, агрегації, дефазифікації;

– побудова УЦП, що передбачає виконання трьох дій: визначення інтервалів розбиття ступенів належності записів групі, вибір типу УЦП, обчислення значень УЦП;

встановлення факту наявності загрози порушення групової анонімності, що передбачає виконання трьох дій: застосування до УЦП ММТТ, вилучення з одержаної множини індексів, що не відповідають викидам, оцінювання адекватності моделі.

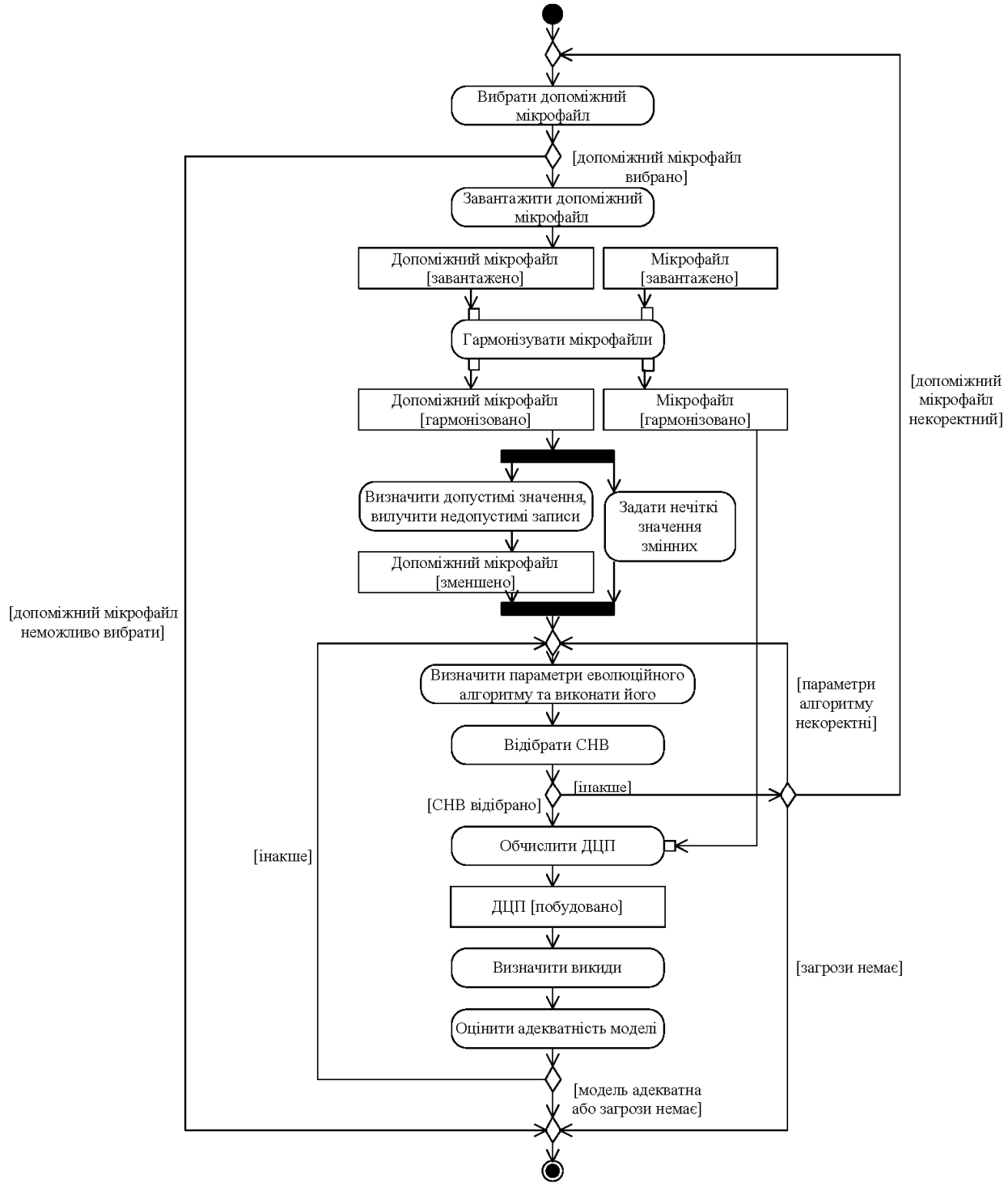

Рис. 4. UML-діаграма активності користувача інформаційної технології на етапі побудови нечіткої моделі групи на основі сторонніх даних

На рис. 5 подано UML-діаграму діяльності користувача на цьому етапі.

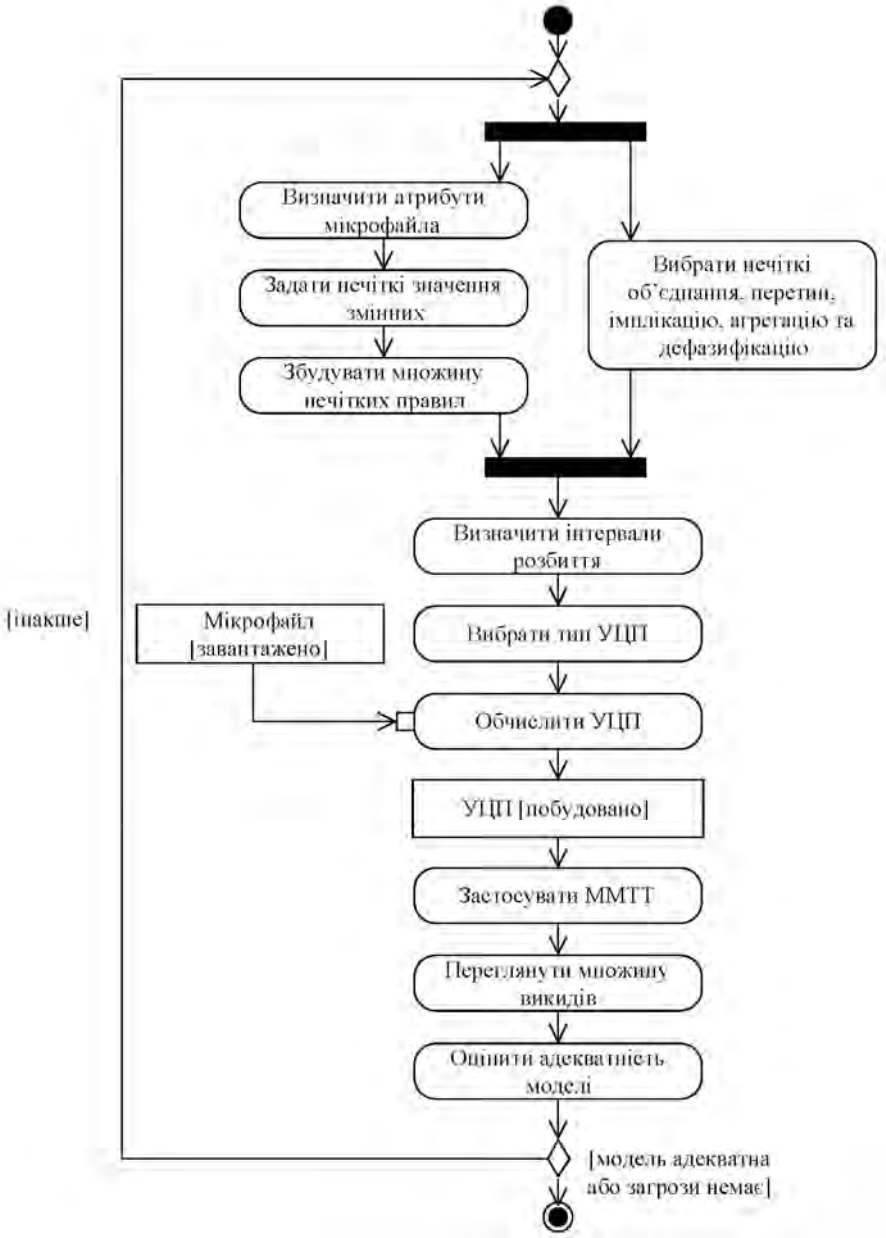

Рис. 5. UML-діаграма активності користувача інформаційної технології на етапі побудови нечіткої моделі групи на основі експертних знань

На етапі *розв'язання ЗЗГА* користувач технології повинен поставити ЗЗГА та застосувати меметичний метод її розв'язання. Для цього потрібно виконати такі операції:

– формування нечітких обмежень на значення ЦП (ДЦП, УЦП), що передбачає виконання трьох дій: вибір K-го найбільшого значення частини ЦП (ДЦП, УЦП), що не відповідає викидам, встановлення порогових значень для відліків, що відповідають викидам, вибір функції належності;

– визначення критеріїв оцінювання якості розв'язку ЗЗГА, що передбачає виконання п'ятьох дій: вибір визначальних атрибутів мікрофайла, встановлення вагових коефіцієнтів для кожного атрибута, встановлення порогів сумісності, спотворень і чутливості;

– виконання меметичного методу розв'язання ЗЗГА, що передбачає виконання трьох дій: визначення параметрів МА, виконання МА, вибір найліпшого розв'язку ЗЗГА; виконання меметичного методу може відбуватися ітеративно;

– модифікація мікрофайла та запис даних модифікованого мікрофайла у вихідний файл.

На рис. 6 подано UML-діаграму діяльності користувача на етапі розв’язання ЗЗГА.

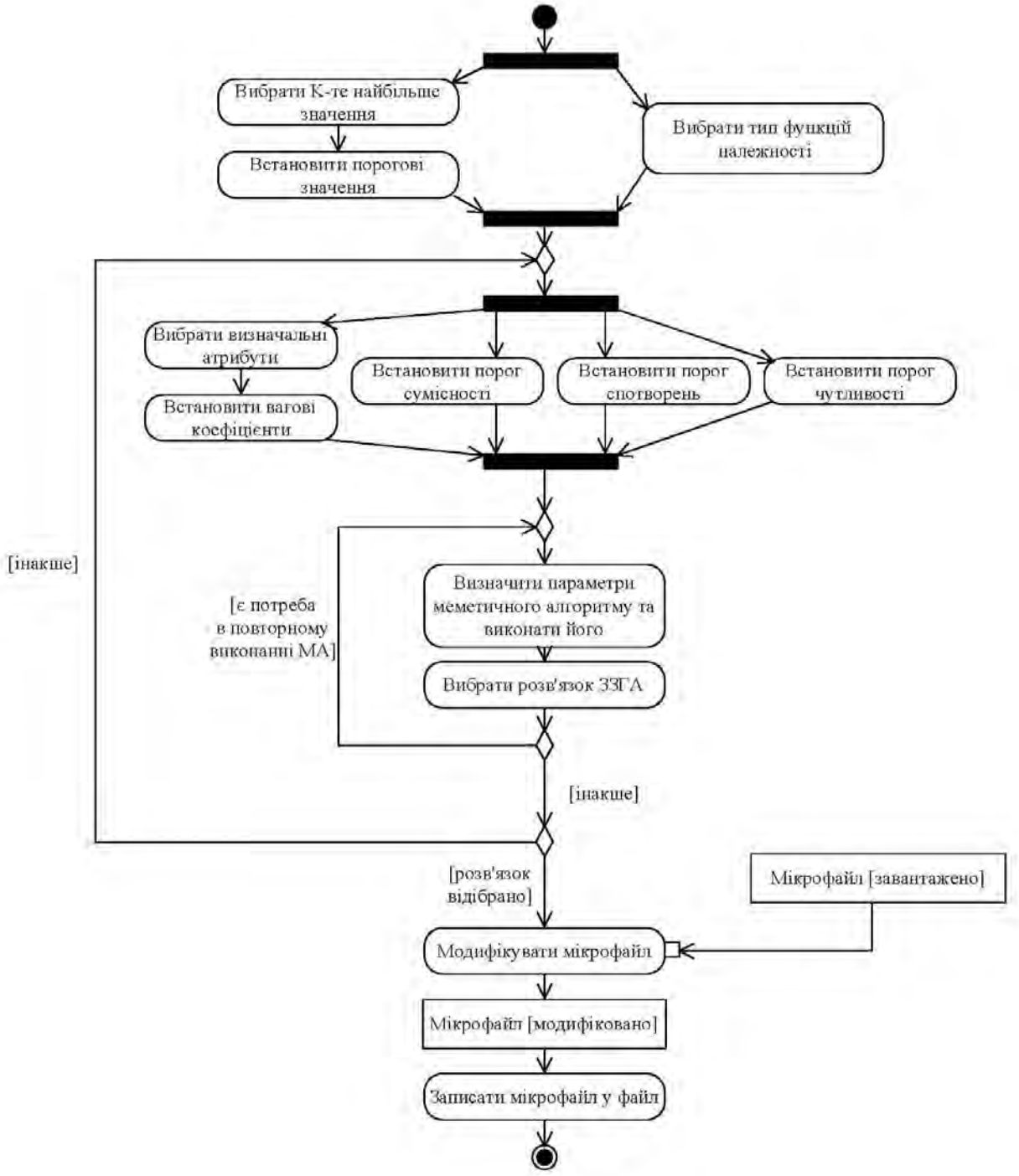

Рис. 6. UML-діаграма активності користувача інформаційної технології на етапі розв’язання задачі забезпечення групової анонімності

На рис. 7 представлено UML-діаграму діяльності користувача під час забезпечення групової анонімності даних, яка охоплює всі етапи технології [21].

На рис. 8 наведено UML-діаграму компонентів інформаційної технології. Кожний компонент реалізує певні операції, описані вище.

Модуль побудови ЦП на основі даних мікрофайла та параметрів групи (сутнісних та параметризувальних атрибутів і значень), одержаних від користувача інформаційної технології за допомогою графічного інтерфейсу, обчислює значення ЦП мікрофайла відносно заданої групи.

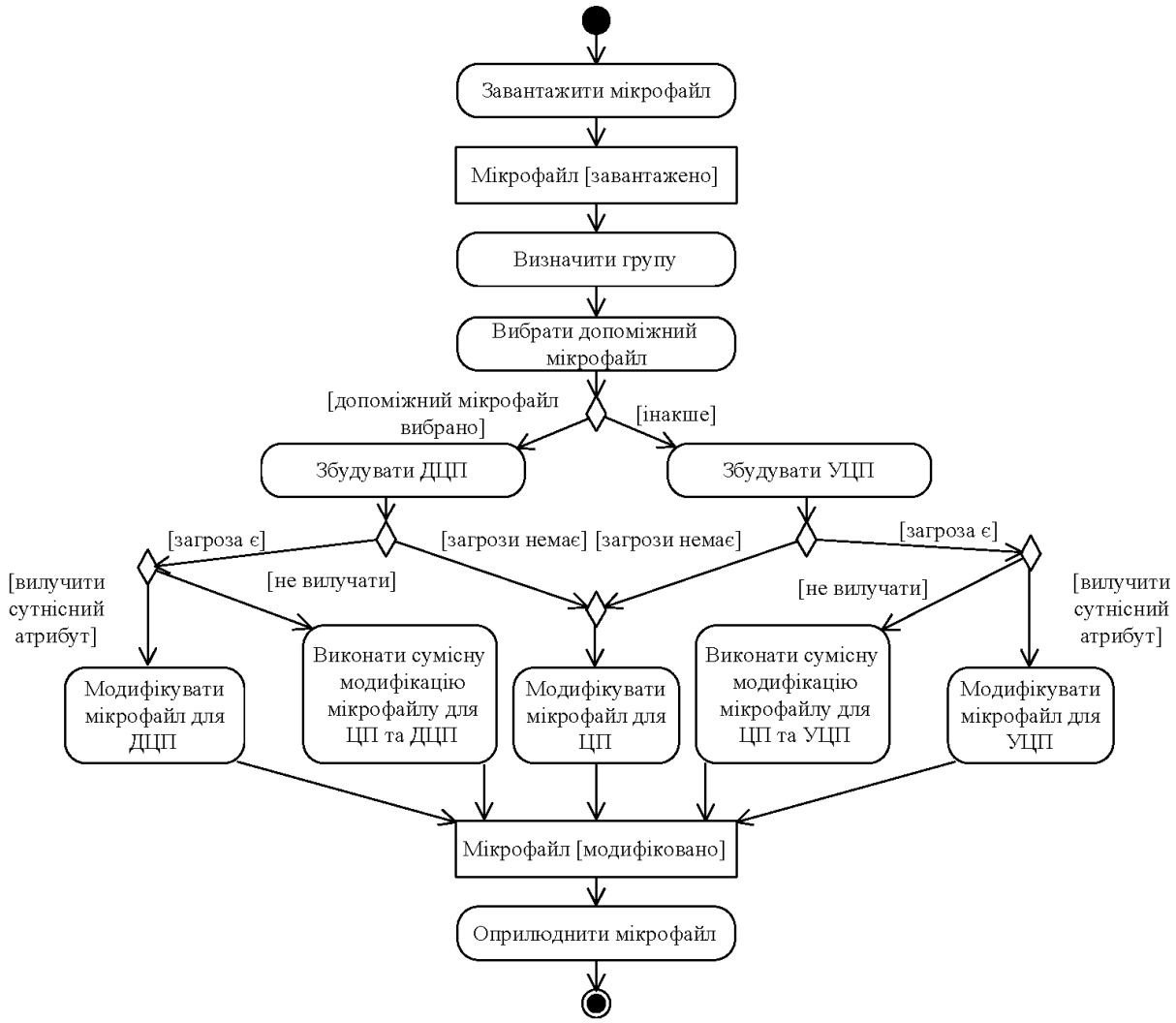

Рис. 7. UML-діаграма діяльності користувача під час забезпечення групової анонімності

Модуль гармонізації мікрофайлів на основі параметрів гармонізації (які атрибути чи їхні значення об'єднувати), одержаних від користувача інформаційної технології за допомогою графічного інтерфейсу, забезпечує гармонізацію основного та допоміжного мікрофайлів. Дані одержаного у такий спосіб гармонізованого допоміжного мікрофайла разом із параметрами ЕА побудови нечіткої моделі групи на основі сторонніх даних, одержаними за допомогою графічного інтерфейсу користувача, використовуються в еволюційному алгоритмі. Результати роботи алгоритму використовуються в нечіткій моделі на основі сторонніх даних, яка, своєю чергою, забезпечує обчислення значень ДЦП гармонізованого мікрофайла.

Модуль побудови УЦП обчислює значення УЦП мікрофайла за допомогою нечіткої моделі на основі експертних знань, параметри СНВ якої (вхідні та вихідні змінні, нечіткі правила, способи реалізації основних нечітких операцій) одержують за допомогою графічного інтерфейсу користувача.

Модуль перевірки адекватності моделі забезпечує визначення викидів у відповідному поданні (ЦП, ДЦП або УЦП). Результати перевірки адекватності відображаються в графічному інтерфейсі.

Меметичний алгоритм на основі параметрів, одержаних за допомогою графічного інтерфейсу користувача, забезпечує розв'язання відповідної ЗЗГА. Модуль модифікації мікрофайла на основі розв'язків ЗЗГА, одержаних за допомогою меметичного алгоритму, забезпечує модифікацію початкового мікрофайла з одержанням модифікованого мікрофайла.

Програмну реалізацію моделей та методів, які є складовою частиною інформаційної технології, виконано мовою програмування *m* у середовищі MATLAB. Вибір цього середовища пояснюється такими міркуваннями:

– вбудована підтримка алгоритмів для роботи з матрицями (якими є, зокрема, мікрофайли);

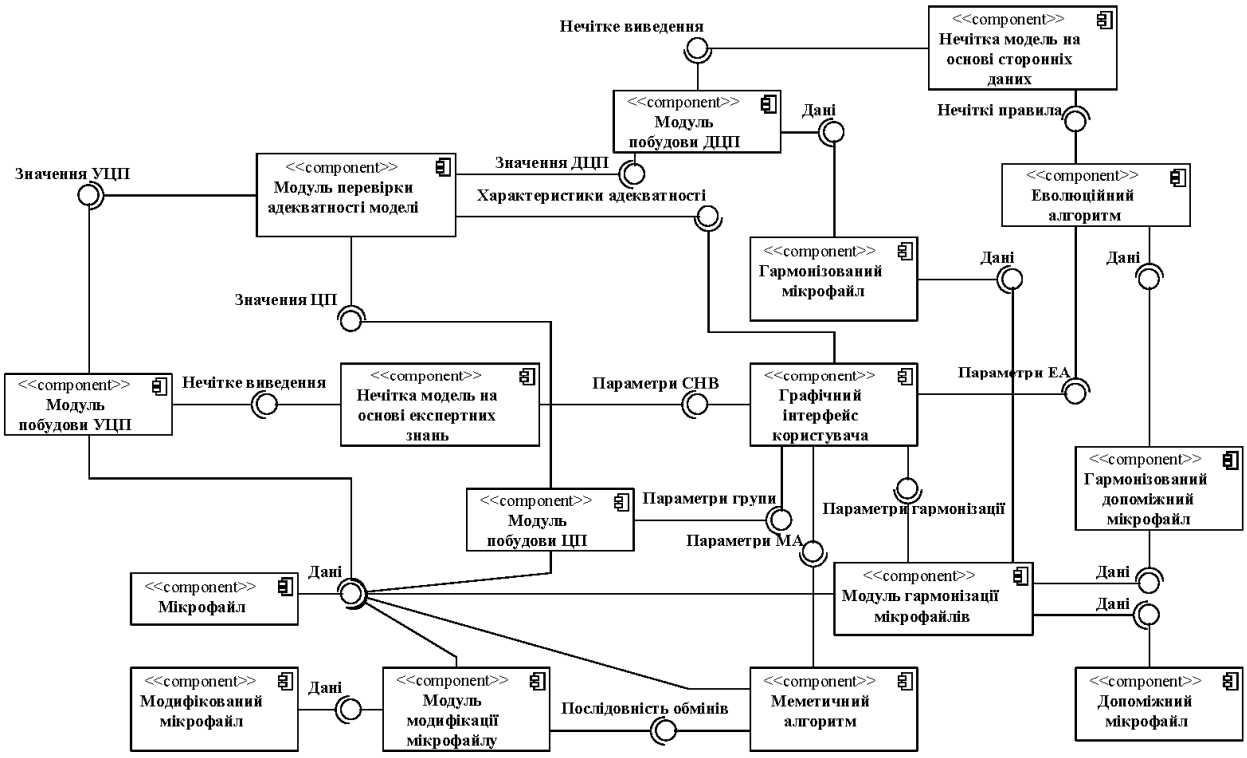

Рис. 8. UML-діаграма компонентів інформаційної технології забезпечення групової анонімності даних

– підтримка паралельних обчислень, які використано під час реалізації ЕА побудови нечіткої моделі групи та МА розв'язання ЗЗГА для пришвидшення обчислення функцій пристосованості особин;

– наявність модуля для роботи з нечіткою логікою, який реалізує метод інтерполяції для модуля нечіткого виведення СНВ типу Мамдані.

Графічний інтерфейс користувача та інтерфейси взаємодії з MATLAB і мікрофайлами реалізовано за допомогою мови програмування Java.

Приклад застосування інформаційної технології

Для ілюстрації застосування інформаційної технології до ЗЗГА, яка ґрунтується на реальних даних, розглянемо задачу маскування територіального розподілу військовослужбовців, що працюють у штаті Флорида, США. Як первинні дані використано файл мікроданих Спостереження за американським суспільством (American Community Survey) 2013 р. [22], який містить 79 284 записи. Вважатимемо, що сутнісний атрибут із мікрофайла не вилучається.

Для визначення групи військовослужбовців, розподілених за місцем їхньої роботи, як сутнісний атрибут було взято атрибут "Професія", як сутнісні значення – значення "551010", "552010", "553010" та "559830" цього атрибута (відповідають військовослужбовцям з різними званнями). Мікрофайл містить атрибути "Місце роботи: штат" і "Місце роботи: область мікроданих публічного використання" (ОМПК). Конкатенація значень цих двох атрибутів дає унікальний код статистичної області, у якій працює респондент. Утворений у результаті атрибут "Місце роботи" є параметризувальним для поставленої задачі.

Кількісний сигнал, що відповідає групі, зображено на рис. 9 суцільною лінією.

Застосувавши ММТТ з параметром $\alpha = 0,01$, можна одержати множину індексів викидів $OUT(\mathbf{q}) = \{1,2,3,4,11,12,15,20,21,24,26,27,29,31,33,34,35,37\}$. Аналізуючи звіт Міністерства оборони США за 2014 р. [23, с. DOD-28–DOD-29], можна зробити висновок, що більшість із індексів, одержаних за допомогою формального застосування ММТТ, не відповідають місцям

розташування військових баз. Далі вважатимемо, що кількісний сигнал має сім викидів, $OUT_e(\mathbf{q}) = \{2,3,11,12,15,24,26\}$:

- викид у 2-й ОМПК відповідає військовим базам Пенема Сіті (Panama City) та Тиндел (Tyndall);
- викид у 3-й ОМПК відповідає базам Кейп Кеневерел (Cape Canaveral) та Петрік (Patrick);
- викид в 11-й ОМПК відповідає базам Джексонвіл (Jacksonville) та Мейпорт (Mayport);
- викид у 12-й ОМПК відповідає базі Пенсакола (Pensacola);
- викид у 15-й ОМПК відповідає базі Мак-Дил (MacDill);
- викид у 24-й ОМПК відповідає базам Говмстед (Homestead) та Кі Вест (Key West);
- викид у 26-й ОМПК відповідає базам Еглін (Eglin) та Герлберт Філд (Hurlburt Field).

Отже, анонімність групи можна забезпечити, зменшивши значення елементів сигналу з індексами 2, 3, 11, 12, 15, 24 та 26. Для цього на $\mathbf{q}^*$ як на 38-арну змінну потрібно накласти нечіткі обмеження:

$$R(X_2): X_2 \text{ is } A_2,$$
$$R(X_3): X_3 \text{ is } A_3,$$
$$R(X_{11}): X_{11} \text{ is } A_{11},$$
$$R(X_{12}): X_{12} \text{ is } A_{12},$$
$$R(X_{15}): X_{15} \text{ is } A_{15},$$
$$R(X_{24}): X_{24} \text{ is } A_{24},$$
$$R(X_{26}): X_{26} \text{ is } A_{26},$$
$$R\left(X_{\bar{J}} \mid q_J^*\right): \sum_{i \in \bar{J}} q_i^* = 306 - \sum_{j \in J} q_j^*,$$

де $J = \{2,3,11,12,15,24,26\}$, $\bar{J} = \{1,4,\ldots,10,13,14,16,\ldots,23,25,27,\ldots,38\}$.

Для кожного з нечітких обмежень вибрано такі функції належності нечітких множин A_i, $i \in J$:

$$\mu_2(x) = ZMF(x,7,21),\ \mu_3(x) = ZMF(x,7,15),\ \mu_{11}(x) = ZMF(x,7,66),$$
$$\mu_{12}(x) = ZMF(x,7,59),\ \mu_{15}(x) = ZMF(x,7,35),\ \mu_{24}(x) = ZMF(x,7,20),$$
$$\mu_{26}(x) = ZMF(x,7,62),$$

де $ZMF(x,a,b) = \begin{cases} 1, & x \le a \\ 1 - 2\left(\dfrac{x-a}{b-a}\right)^2, & a \le x \le \dfrac{a+b}{2} \\ 2\left(\dfrac{x-b}{b-a}\right)^2, & \dfrac{a+b}{2} \le x \le b \\ 0, & x \ge b \end{cases}$.

Порогові значення підібрано з міркувань зменшення значень відповідних елементів сигналу до рівня, зіставного з величиною найменшого викиду.

Для мінімізації обсягу внесених у мікрофайл спотворень як визначальні атрибути взято “Вік”, “Рівень освіти [узагальнений]”, “Стать”, “Раса [узагальнена]”, “Кількість робочих годин на тиждень”, “Іспанське походження [узагальнене]”, “Сімейний стан”, “Спосіб добирання на роботу”, “Час виходу на роботу”, “Час на дорогу на роботу”, “Кількість робочих тижнів на рік (інтервал)”, “Сукупний дохід” та “Володіння англійською”. Кожний атрибут вважався категорійним. Для спрощення інтерпретації метрики вибрано такі параметри (4): $\gamma_k = 1 \ \ \forall k = \overline{1,13}$, $\chi_1 = 1$, $\chi_2 = 0$. Метрика (4) в цьому разі показує кількість значень атрибутів, які потрібно модифікувати за один обмін записів між підмікрофайлами.

Функція пристосованості МА набуває такого вигляду:

$$f^{ex}(U)=\frac{3614-\sum_{i=1}^{Q}\sum_{k=1}^{13}\operatorname{sign}\left|\mathbf{M}_{u_{i1}}(u_{i2},W_k)-\mathbf{M}_{u_{i3}}(u_{i4},W_k)\right|}{3614}\times$$

$$\times\prod_{j\in J}\mu_j\left(q_j^*(U)\right)\cdot\frac{1}{1+e^{\frac{1}{2}(Q_U-240)}},$$

де перший множник виражає якість розв'язку з погляду мінімізації спотворень, другий – якість розв'язку з погляду сумісності з накладеними нечіткими обмеженнями, третій – штрафний терм, уведений з метою упередження необмеженого зростання особин в МА; U – особина в алгоритмі, що являє собою [7] матрицю розмірності $Q\times 4$, з елементами u_{ij}; $q_j^*(U)$ – значення модифікованого кількісного сигналу, утвореного виконанням обмінів записів, заданих особиною U; 3614 – максимально можливе значення сумарної метрики (4) для цієї задачі; W_k – k-й визначальний атрибут, $k=\overline{1,13}$; $\mathbf{M}_j(i,W_k)$ – оператор, який повертає значення атрибута W_k i-го запису підмікрофайла $\mathbf{M}_j$.

Як оператори мутації та рекомбінації вибрано оператори, описані в [4]. Як оператор відбору вибрано оператор турнірного відбору [24]. Вибрано розмір турніру 5. Популяцію в МА ініціалізовано випадковим генеруванням матриць із різною кількістю рядків. Інші параметри алгоритму вибрано так: розмір популяції $\mu=100$, кількість особин, які замінюють на кожному поколінні $\lambda=40$, імовірність рекомбінації $p_c=1$, імовірність мутації $p_{m_1}=p_{m_2}=p_{m_3}=p_{m_4}=0,001$, параметр застосування локального пошуку $p_{mem}=0,75$. Вибір цих параметрів є типовим для задач відповідного класу. Для упередження передчасної збіжності алгоритму ймовірність мутації збільшувалася вдесятеро щоразу, коли середньоквадратичне відхилення пристосованостей особин у популяції ставало менше за 0,03.

Параметри ЗЗГА для оцінювання якості розв'язку було визначено так:

- значення порога сумісності – $\alpha_{comp}=0,5$;
- значення порога чутливості – $K_{out}=0$;
- значення порога спотворень – $K_{dist}=0,25$.

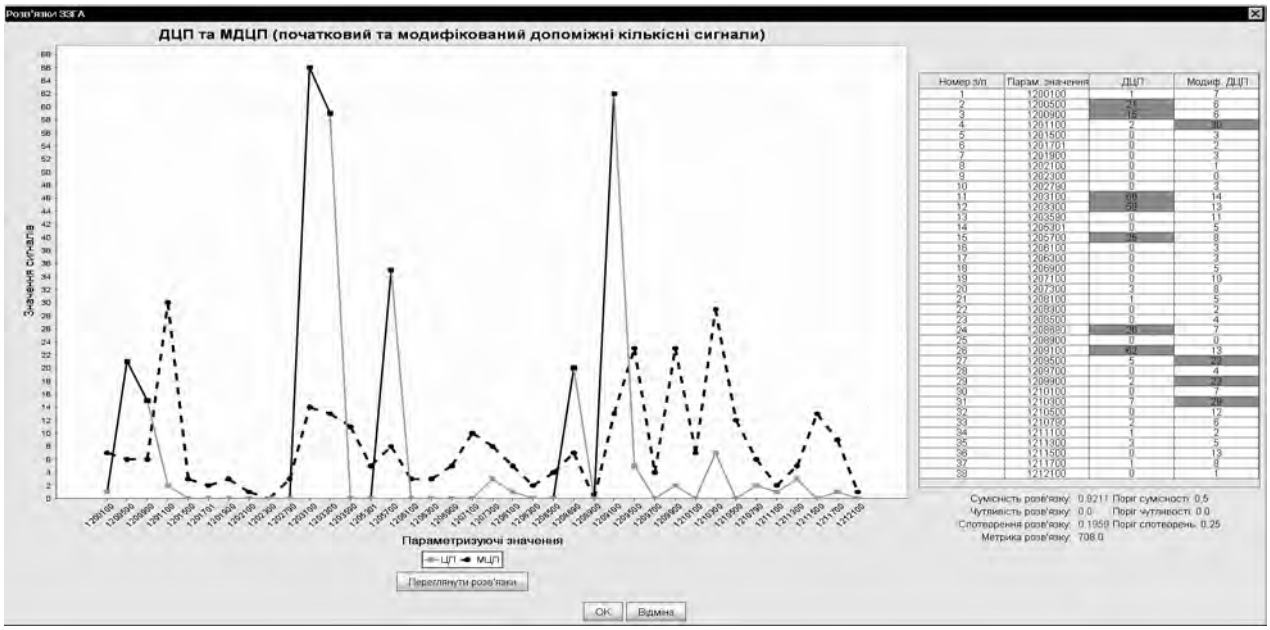

Рис. 9. Початковий (суцільна лінія) та модифікований (штрихова лінія) кількісний сигнал зі значенням метрики 708

Виконано 30 окремих запусків МА. Кожного запуску алгоритм припиняв свою роботу після генерації 1000 популяцій. Серед 3000 індивідів, отриманих за результатами застосування МА, 2580 штук є допустимими розв'язками ЗЗГА. Розв'язок із найменшою сумарною метрикою (4) подано на рис. 9. Він є допустимим, зокрема, тому, що застосування ММТТ до нього дає множину $OUT\left(\mathbf{q}^*\right)=\{4,27,29,31\}$, тобто в модифікованому кількісному сигналі немає викидів, що відповідають викидам початкового сигналу. Середнє значення сумарної метрики (4) по всіх допустимих розв'язках дорівнює 782,2209. Інакше кажучи, для забезпечення групової анонімності достатньо змінити не більше за $\frac{782,2209}{13\cdot 79\,284}\approx 0,08\ \%$ значень атрибутів мікрофайла.

Висновки та перспективи подальших наукових розвідок

У статті задачу забезпечення групової анонімності даних зведено до узагальненої задачі пошуку потоку мінімальної вартості в мережі, на архітектуру якої накладено певні нечіткі обмеження. Запропоновано нову інформаційну технологію забезпечення анонімності даних про групи, відносно яких існує загроза її порушення шляхом аналізу комбінацій значень базових атрибутів мікрофайла. На відміну від відомих засобів забезпечення анонімності, ця технологія враховує вплив значень базових атрибутів і не потребує участі експерта на етапі оцінювання якості розв'язку ЗЗГА.

Експеримент з анонімізації реальних даних статистичного спостереження за американським суспільством у 2013 р. показав, що за умови застосування цієї технології обсяг спотворення даних не перевищує 0,08 % від загальної кількості значень атрибутів.

1. Fung B. Privacy-preserving data publishing: a survey of recent developments / B. Fung, K. Wang, R. Chen, P. Yu // ACM Computing Surveys. – 2010. – 42(4). – P. 1–53. 2. Chertov O. Group Anonymity / O. Chertov, D. Tavrov // Information Processing and Management of Uncertainty in Knowledge-Based Systems. Applications [ed. E. Huellermeier, R. Kruse, F. Hoffmann]. – Berlin, Heidelberg : Springer-Verlag, 2010. – P. 592–601. 3. Sweeney L. Computational Disclosure Control: A Primer on Data Privacy Protection / L. Sweeney // Ph.D. Thesis. – Massachusetts Institute of Technology, Cambridge, 2001. – 216 p. 4. Chertov O. Microfiles as a Potential Source of Confidential Information Leakage / O. Chertov, D. Tavrov // Intelligent Methods for Cyber Warfare [ed. R. R. Yager, M. Z. Reformat, N. Alajlan]. – Springer International Publishing Switzerland, 2015. – P. 87–114. – (Studies in Computational Intelligence, vol. 563). 5. Chertov O. Memetic Algorithm for Solving the Task of Providing Group Anonymity / O. Chertov, D. Tavrov // Advance Trends in Soft Computing [ed. M. Jamshidi, V. Kreinovich, J. Kacprzyk]. – Springer International Publishing Switzerland, 2014. – P. 281–292. – (Studies in Fuzziness and Soft Computing, vol. 312). 6. Ahuja R. K. Network Flows. Theory, Algorithms, and Applications / R. K. Ahuja, T. L. Magnanti, J. B. Orlin. – Upper Saddle River : Prentice Hall, 1993. – 864 p. 7. Чертов О. Р. Меметичний алгоритм із нечіткими обмеженнями для розв'язання задачі забезпечення групової анонімності / О. Р. Чертов, Д. Ю. Тавров // Інформаційна безпека. – 2013. – № 4 (12). – С. 135–144. 8. Sweeney L. Replacing personally-identifying information in medical records, the Scrub system / L. Sweeney // Proceedings, Journal of the American Medical Informatics Association [ed. J. J. Cimino]. – Washington, DC : Hanley & Belfus, 1996. – P. 333–337. 9. Sweeney L. Datafly: a system for providing anonymity in medical data / L. Sweeney // Proceedings of the IFIP TC11 WG11. 3 Eleventh International Conference on Database Securty XI: Status and Prospects. – London : Chapman & Hall, Ltd., 1998. – P. 356–381. 10. μ-ARGUS version 5.1. User's Manual [Electronic Resource] / [A. Hundepool, P.-P. de Wolf, J. Bakker, A. Reedijk, L. Franconi et al.]. – 2014. – 88 p. – Mode of access: http://neon.vb.cbs.nl/casc/Software/MUmanual5.1.pdf. 11. Чертов О. Р. Моделі, інформаційні технології та архітектура систем обробки демографічної інформації: дис. … доктора техн. наук : 05.13.06 / Чертов Олег Романович. – К., 2014. – 383 с. 12. Test Uncertainty : PTC 10.1-2013. – [Чинний від 2013-01-01]. – ASME, 2013. – 112 p. 13. Student. The probable error of a mean / Student // Biometrika. – 1908. – Vol. 6,

No. 1. – P. 1–25. 14. Zadeh L. A. The Concept of a Linguistic Variable and its Application to Approximate Reasoning / L. A. Zadeh // Information Sciences. – 1975. – 8. – P. 199–249. 15. Чертов О. Р. Эволюционный алгоритм построения нечеткой модели группы с целью нарушения ее анонимности / О. Р. Чертов, Д. Ю. Тавров // Международная научная конференция имени Т. А. Таран "Интеллектуальный анализ информации" ИАИ-2015, Киев, 20–22 мая 2015 г. : сб. тр. / гл. ред. С. В. Сирота. – К. : Просвіта, 2015. – С. 272–280. 16. Group Methods of Data Processing / [Chertov O., Tavrov D., Pavlov D. et al.] ; ed. O. Chertov. – Raleigh : Lulu.com, 2010. – 156 p. 17. Автоматизовані системи. Терміни та визначення : ДСТУ 2226-93. – К. : Держстандарт України, 1994. – 94 с. 18. Павлов А. А. Информационные технологии и алгоритмизация в управлении / А. А. Павлов, С. Ф. Теленик. – К. : Техніка, 2002. – 344 с. 19. Информатика : учебник / [Макарова Н. В., Матвеев Л. А., Бройдо В. Л. и др.] ; под ред. Н. В. Макаровой. – [3-е изд., перераб.]. – М. : Финансы и статистика, 2009. – 768 с. 20. Rumbaugh J. The Unified Modeling Language Reference Manual / J. Rumbaugh, I. Jacobson, G. Booch. – [2nd ed.]. – Addison-Wesley, 2004. – 721 p. 21. Тавров Д. Ю. Схема інформаційної технології забезпечення групової анонімності даних / Д. Ю. Тавров // Системний аналіз та інформаційні технології : матеріали 17-ї Міжнародної науково-технічної конференції SAIT 2015, Київ, 22–25 червня 2015 р. / ННК "ІПСА" НТУУ "КПІ". – К. : ННК "ІПСА" НТУУ "КПІ", 2015. – С. 290–291. 22. Integrated Public Use Microdata Series, Version 5.0 [Machine-readable database] [Electronic resource] / S. Ruggles, J. T. Alexander, K. Genadek, R. Goeken, M. B. Schroeder, M. Sobek. – Minneapolis : University of Minnesota, 2010. – Mode of access: https://usa.ipums.org/usa/. 23. Base Structure Report (A Summary of the Real Property Inventory) Fiscal Year 2014 Baseline [Electronic resource] / Office of the Deputy under Secretary of Defense. – 2013. – Mode of access: http://www.acq.osd.mil/ie/download/bsr/Base%20Structure%20Report%20FY14.pdf. 24. Baker J. E. Reducing bias and inefficiency in the selection algorithm / J. E. Baker // Proceedings of the 2nd International Conference on Genetic Algorithms and Their Applications [ed. J. J. Grefenstette]. – Hillsdale, New Jersey : Lawrence Erlbaum, 1987. – P. 14–21.