DOI: 10.18372/2410-7840.21.13765
УДК 004.056.53(045)

МЕТОД СТРУКТУРИЗАЦІЇ КОДОВИХ КОНСТРУКЦІЙ НА ОСНОВІ ГАУСІВСЬКОЇ ЗМІШАНОЇ МОДЕЛІ ТА ВИДІЛЕННЯ ЇЇ КОМПОНЕНТ

Олексій Голубничий

Методи та підходи лінійного криптоаналізу криптографічних алгоритмів спрямовані на аналіз та виявлення взаємозв'язків між елементами відкритого тексту, шифротексту та ключа. У випадку лінійного криптоаналізу псевдовипадкових послідовностей та сигнально-кодових конструкцій, які побудовані на їх основі, аналізу та виявленню підлягають взаємозв'язки між елементами цих послідовностей та сигнально-кодових конструкцій, а також між їх внутрішніми структурами, їх складовими в утворюваних системах кодових конструкцій тощо. Ефективність реалізації лінійного криптоаналізу на різних його етапах може бути підвищена при структуризації (виявлення внутрішніх структур та взаємозв'язків між ними) досліджуваних кодових конструкцій, щодо яких у сторони здійснення криптоаналізу відсутня будь-яка апріорна інформація про їх структуру, або кодових конструкцій, які апріорі можуть вважатися такими, що мають стохастичну природу їх утворення. У статті запропоновано метод структуризації кодових конструкцій з апріорі невідомою структурою на основі аналізу кореляційних зв'язків між кодовими конструкціями, які представляються у цьому методі гаусівською змішаною моделлю з подальшим виділенням її компонент та кластеризацією досліджуваних кодових конструкцій з використанням обґрунтованої у статті параметрично-критеріальної модифікації ЕМ-алгоритму з видаленням компонент. Метод дозволяє виокремлювати групи кодових конструкції з взаємопов'язаними структурами і далі виділяти ці взаємопов'язані структури у явному вигляді, в чому може полягати розв'язання ряду задач лінійного криптоаналізу, які пов'язані з виявленням структур та взаємозв'язків між ними. Показано приклад реалізації запропонованого методу для структуризації бінарних псевдовипадкових послідовностей Баркера, які використовуються як сигнально-кодові конструкції у широкосмугових системах передавання інформації та для яких з літературних джерел відомо, що вони були синтезовані методом напрямленого перебору і тому мають стохастичний апріорі неструктурований характер.

Ключові слова: *лінійний криптоаналіз, структуризація кодових конструкцій, виявлення структур, виявлення взаємозв'язків, гаусівська змішана модель, ЕМ-алгоритм, псевдовипадкові послідовності, машинне навчання.*

Вступ

Методи та підходи лінійного криптоаналізу криптографічних алгоритмів спрямовані на аналіз та виявлення взаємозв'язків між елементами відкритого тексту, шифротексту та ключа [1, 2]. У випадку лінійного криптоаналізу псевдовипадкових послідовностей (ПВП) та кодових конструкцій, які побудовані на їх основі, аналізу та виявленню підлягають взаємозв'язки між елементами цих послідовностей та кодових конструкцій, а також між їх внутрішніми структурами, їх складовими в утворюваних системах кодових конструкцій тощо. Сучасні дослідження криптографічних властивостей криптографічних алгоритмів, як правило, містять у своєму складі аналіз криптографічної стійкості криптографічного алгоритму до лінійного та інших видів криптоаналізу. До таких досліджень належать, наприклад, дослідження [3], які містять напрям оцінки та обґрунтування стійкості прийнятого у 2015 р. в Україні національного стандарту шифрування ДСТУ 7624:2014, який визначає блоковий шифр "Калина" [4, 5], до алгебраїчних та статистичних методів різницевого та лінійного криптоаналізу, алгебраїчних атак, заснованих на гомоморфізмах. У той же час відомо, що стійкість до лінійного криптоаналізу не входила у число критеріїв при проектуванні, наприклад, відомого криптографічного алгоритму DES [1, с. 294]. Важливим є також аналіз криптографічної стійкості ПВП, які використовуються для захисту інформації від несанкціонованого доступу у методах гамування, потокових шифрах, стеганографічних методах з розширенням спектру тощо [1, 2]. Певні структурні взаємозв'язки між елементами ПВП є наслідком детермінованих правил синтезу, які використовуються у відповідних генераторах ПВП. Використання таких детермінованих правил синтезу з одного боку дає можливість відтворювати ПВП у різних складових частинах інформаційно-телекомунікаційної системи для санкціонованого розшифровування повідомлень, наприклад під час складання по схемі XOR при гамуванні на приймальному та передавальному боці, а з іншого – вносить певну структурованість у елементи ПВП і, як наслідок, шифротексту, який отримують при використанні цих ПВП, або інших кодових структур, наприклад S-блоків блокового шифру [6, 7]. Підвищенню стійкості ПВП до лінійного криптоаналізу [8], а також методам синтезу різних гібридних крипто-кодових конструкцій [9], складних дискретних сигналів на основі кодових конструкцій з певними кореляційними властивостями [10] присвячується значна увага зокрема через те, що статистичні характеристики ПВП або кодових

конструкцій можуть бути суттєвим чином пов'язані з їх структурованістю, що впливає на рівень криптографічної стійкості алгоритму шифрування або на стійкість до атак стеганографічного методу захисту інформації, в якому використовуються такі ПВП та кодові конструкції. Для перевірки рівня стохастичності та структурованості ПВП часто використовується набір з 15-и статистичних тестів *NIST* [11], які виконують перевірку ПВП за такими критеріями: частотний побітовий тест (*Frequency Monobit Test*), частотний блоковий тест (*Frequency Test within a Block*), тест на послідовність однакових бітів (*Runs Test*), тест на найдовшу послідовність одиниць у блоці (*Test for the Longest Run of Ones in a Block*), тест рангів матриць (*Binary Matrix Rank Test*), спектральний тест на основі дискретного перетворення Фур'є (*Discrete Fourier Transform (Spectral) Test*), тест на співпадіння структур (шаблонів), які не перекриваються (*Non-overlapping Template Matching Test*), тест на співпадіння структур (шаблонів), які перекриваються (*Overlapping Template Matching Test*), універсальний статистичний тест Маурера (*Maurer's "Universal Statistical" Test*), тест лінійної складності (*Linear Complexity Test*), тест періодичності (*Serial Test*), ентропійний тест (*Approximate Entropy Test*), тест кумулятивних сум (*Cumulative Sums (Cusum) Test*) та дві варіації тестів випадкових відхилень (*Random Excursions Test* та *Random Excursions Variant Test*).

В рамках специфіки дослідження, яке викладається у цій статті, зазначимо, що критерії у значній кількості вищезазначених тестів *NIST* можна пов'язати з кореляційними властивостями ПВП, а саме з автокореляційною функцією ПВП або кодової конструкції. Відомо, що стохастичні властивості ПВП повинні наближатися до характеристик гаусівського білого шуму, оскільки його диференціальна ентропія є максимальною для неперервних функцій часу з обмеженою середньою потужністю [12, с. 63], внаслідок чого такі ПВП з інформаційної точки зору вноситимуть найбільшу невизначеність до шифротексту під час криптографічних перетворень. У той же час білий шум має автокореляційну функцію у вигляді δ-функції Дірака через статистично непов'язані (некорельовані) значення такого процесу, що у свою чергу забезпечує його рівномірний спектр [13, с. 104]. Вимоги саме до таких властивостей автокореляційної функції опосередковано висуваються до ПВП в тестах *NIST* і підлягають перевірці, наприклад, у спектральному тесті, в якому відбувається перевірка ступеня рівномірності спектру ПВП, або у тесті на найдовшу послідовність одиниць у блоці (ПВП з автокореляційною функцією, яка наближена за формою до δ-функції Дірака, не може мати блоків зі значною кількістю одиниць у своїй структурі, оскільки це призвело б до викидів значень у бічних пелюстках автокореляційної функції і спричинило суттєві відхилення від форми δ-функції Дірака).

Актуальність дослідження полягає в тому, що у галузі технічного захисту інформації продовжується розробка нових методів синтезу ПВП, гібридних крипто-кодових конструкцій, складних дискретних сигналів на основі кодових конструкцій з певними кореляційними властивостями тощо для забезпечення необхідного рівня криптографічної стійкості алгоритмів шифрування або стійкості до атак стеганографічних методів захисту інформації. При цьому одним з основних інструментів для тестування ПВП та кодових конструкцій є набір статистичних тестів *NIST* [11], які за своїми критеріями дають чисельні характеристики та оцінки відповідності рівня криптографічної стійкості досліджуваних ПВП або кодових конструкцій. Метод структуризації, який пропонується у цій статті, може виявляти у ПВП та кодових конструкціях внутрішні структури та взаємозв'язки між ними, якщо вони у них присутні, що дозволяє здійснювати більш детальний, ніж при інтегрованих чисельних характеристиках та оцінках, аналіз на предмет наявності структурованості ПВП або кодових конструкцій з виділенням цієї структурованості в явному вигляді. Результати такого аналізу можуть бути корисними при розробці криптографічних методів захисту інформації, методів синтезу ПВП та кодових конструкцій, а також для розв'язання задач лінійного криптоаналізу.

Наукова новизна дослідження полягає в тому, що запропоновано новий метод структуризації кодових конструкцій з апріорі невідомою для аналітика структурою на основі аналізу кореляційних зв'язків між кодовими конструкціями, які представляються у цьому методі гаусівською змішаною моделлю з подальшим виділенням її компонент та кластеризацією досліджуваних кодових конструкцій з використанням параметрично-критеріальної модифікації ЕМ-алгоритму з видаленням компонент. Особливістю методу є те, що він дозволяє виокремлювати групи кодових конструкції з взаємопов'язаними структурами і далі виділяти ці взаємопов'язані структури у явному вигляді.

Формулювання цілей статті

Метою статті є математична формалізація, опис та реалізація на прикладі процедур нового методу структуризації кодових конструкцій з апріорі невідомою структурою, який для статистичного аналізу використовує гаусівську змішану модель та параметрично-критеріальну модифікацію ЕМ-алгоритму з видаленням компонент для виділення компонент гаусівської змішаної моделі.

Виклад основного матеріалу дослідження

Метод структуризації кодових конструкцій, який пропонується у цій статті, викладено поетапно у такій послідовності: 1) задання вхідних даних; 2) кореляційний аналіз кодових конструкцій; 3) аналіз кореляційних зв'язків між кодовими конструкціями на основі гаусівської змішаної моделі, виділення її компонент з використанням параметрично-критеріальної модифікації ЕМ-алгоритму та структуризація виокремлених кодових конструкцій.

Етап 1. Задання вхідних даних.

В якості кодових конструкцій, які підлягають структуризації (виявленню внутрішніх структур та взаємозв'язків між ними), у статті для прикладу було обрано бінарні послідовності Баркера [13, с. 108] з таких причин:

1) бінарні послідовності Баркера були синтезовані методом напрямленого перебору [14, с. 23] і тому можуть вважатися кодовими конструкціями з апріорі стохастичним неструктурованим характером;

2) бінарні послідовності Баркера мають найменший можливий рівень бічних пелюсток автокореляційної функції серед усіх можливих бінарних послідовностей (опосередкований зв'язок між автокореляційною функцією кодової конструкції, її статистичними характеристиками та рядом критеріїв *NIST* було проаналізовано у вступі);

3) бінарні послідовності Баркера можуть мати як однакову, так і різну довжину, що можна використати для ілюстрації можливостей методу, який пропонується, при аналізі систем кодових конструкцій різної довжини;

4) відомі бінарні послідовності Баркера обмежені за довжиною і кількістю, що дозволяє застосувати метод, який пропонується, для структуризації всіх відомих таких послідовностей та проілюструвати можливості методу для аналізу структур цілого взятого для аналізу типу ПВП.

Особливістю кодових конструкцій, що досліджуються на предмет їх структурованості, є наявність інверсно-ізоморфних структур. Одна й та сама кодова конструкція може бути використана у формі її запису "зліва направо", "справа наліво", "зліва направо з інверсією", "справа наліво з інверсією", утворюючи чотири інверсно-ізоморфні структури однієї і тієї самої кодової конструкції. При цьому різні інверсно-ізоморфні структури мають одну й ту саму форму автокореляційної функції. Зазначимо, що метод структуризації, який пропонується, є чутливим до різних інверсно-ізоморфних структур. У табл. 1 наведено усі відомі бінарні послідовності Баркера довжини $D = 2, 3, 4, 5, 7, 11, 13$ (для $D = 2$ та $D = 4$ існує дві різні послідовності), а також їх можливі інверсно-ізоморфні структури.

Таблиця 1

Відомі бінарні послідовності Баркера та їх інверсно-ізоморфні структури

D	ПВП			
2	1; -1		-1; 1 **	
2	1; 1		-1; -1 **	
3	1; 1; -1 *	-1; -1; 1	-1; 1; 1 **	1; -1; -1
4	1; 1; 1; -1 *	-1; -1; -1; 1	-1; 1; 1; 1 **	1; -1; -1; -1
4	1; 1; -1; 1 *	-1; -1; 1; -1	1; -1; 1; 1	-1; 1; -1; -1 **
5	1; 1; 1; -1; 1 *	-1; -1; -1; 1; -1	1; -1; 1; 1; 1	-1; 1; -1; -1; -1 **
7	1; 1; 1; -1; -1; 1; -1 *	-1; -1; -1; 1; 1; -1; 1	-1; 1; -1; -1; 1; 1; 1 **	1; -1; 1; 1; -1; -1; -1
11	1; 1; 1; -1; -1; -1; 1; -1; -1; 1; -1 *		-1; -1; -1; 1; 1; 1; -1; 1; 1; -1; 1	
	-1; 1; -1; -1; 1; -1; -1; -1; 1; 1; 1 **		1; -1; 1; 1; -1; 1; 1; 1; -1; -1; -1	
13	1; 1; 1; 1; 1; -1; -1; 1; 1; -1; 1; -1; 1 *		-1; -1; -1; -1; -1; 1; 1; -1; -1; 1; -1; 1; -1	
	1; -1; 1; -1; 1; 1; -1; -1; 1; 1; 1; 1; 1		-1; 1; -1; 1; -1; -1; 1; 1; -1; -1; -1; -1; -1 **	

* структури послідовностей, які наведено у [2, с. 108]

** структури послідовностей, які забезпечили можливість здійснення структуризації у дослідженні

Окремо у табл. 1 виділено ті інверсно-ізоморфні структури послідовностей, які забезпечили можливість здійснення структуризації у дослідженні. Вибір тієї чи іншої інверсно-ізоморфної структури для аналізу при використанні методу, який пропонується, характеризується певною апріорною невизначеністю, а також емпіричним та комбінаторним характером.

Таким чином, в якості вхідних даних використовуватимемо такі кодові конструкції, які досліджуються на предмет виявлення їх внутрішніх структур та взаємозв'язків між ними:

$$\mathbf{X}_1=\{-1;\ 1\},\ \mathbf{X}_2=\{-1;-1\},\ \mathbf{X}_3=\{-1;\ 1;\ 1\},$$
$$\mathbf{X}_4=\{-1;\ 1;\ 1;\ 1\},\ \mathbf{X}_5=\{-1;\ 1;-1;-1\},$$
$$\mathbf{X}_6=\{-1;\ 1;-1;-1;-1\},$$
$$\mathbf{X}_7=\{-1;\ 1;-1;-1;\ 1;\ 1;\ 1\},$$
$$\mathbf{X}_8=\{-1;\ 1;-1;-1;\ 1;-1;-1;-1;\ 1;\ 1;\ 1\},$$
$$\mathbf{X}_9=\{-1;\ 1;-1;\ 1;-1;-1;\ 1;\ 1;-1;-1;-1;-1;-1\}.$$

Етап 2. Кореляційний аналіз кодових конструкцій.

Кореляційний аналіз кодових конструкцій у даному дослідженні полягає у аналізі взаємнокореляційних зв'язків між ними. Кореляційні зв'язки мають таку змістовну значимість з точки зору структуризації кодових конструкцій: якщо кореляція між кодовими конструкціями відсутня (наприклад, система кодових конструкцій є ортогональною), то вони є лінійно-незалежними структурами [15, с. 352]; якщо ж існує певна кореляція між кодовими конструкціями, то структура окремо взятої кодової конструкції може бути частково представлена через сукупність структур інших кодових конструкцій [16] (тривіальним випадком є значення нормованої взаємнокореляційної функції, що дорівнює ± 1, коли структура однієї кодової конструкції повністю співпадає з структурою іншої кодової конструкції з точністю до інверсії).

Особливістю кореляційного аналізу в даному дослідженні також є те, що кореляційні зв'язки визначаються для пар бінарних послідовностей Баркера різних довжин таким чином, що послідовності утворюють сигнально-кодові конструкції $x_m(t),\ \ m=\overline{1,M}$ ($M=9$ – кількість кодових конструкцій у системі, що підлягає кореляційному аналізу відповідно до табл. 1), які існують на одному й тому самому інтервалі часу T, що формалізовано виразом (1):

$$x_m(t)=\sum_{i=1}^{D_m} x_i\left\{H\left[t-(i-1)\tau_m\right]-H\left[t-i\tau_m\right]\right\}, \quad (1)$$

де x_i, $i=\overline{1,D_m}$, – елементи послідовності довжини D_m, що утворюють $x_m(t)$, $H(t)$ – одинична ступінчата функція, $\tau_m=T/D_m$ – тривалість елемента послідовності.

Сигнально-кодові конструкції $x_m(t)$, $m=\overline{1,9}$, які побудовано на основі послідовностей $\mathbf{X}_m$ та для яких обчислюється значення взаємної кореляції, показані на рис. 1.

Кореляційний аналіз полягає у визначенні взаємнокореляційних зв'язків $R_{u,v}=\frac{1}{T}\int\limits_{t\in T} x_u(t)x_v(t)dt$, $u=\overline{1,9}$, $v=\overline{1,9}$, у системі сигнально-кодових конструкцій $x_m(t)$, $m=\overline{1,9}$. Результатом такого кореляційного аналізу є матриця коефіцієнтів взаємної кореляції (2):

$$\mathbf{R}=\begin{pmatrix}
1 & 0 & 0,667 & 0,5 & -0,5 & -0,4 & 0,571 & 0,182 & -0,154\\
0 & 1 & -0,333 & -0,5 & 0,5 & 0,6 & -0,143 & 0,091 & 0,385\\
0,667 & -0,333 & 1 & 0,833 & -0,167 & -0,467 & 0,238 & 0,212 & -0,333\\
0,5 & -0,5 & 0,833 & 1 & 0 & -0,3 & 0,214 & 0,045 & -0,269\\
-0,5 & 0,5 & -0,167 & 0 & 1 & 0,7 & -0,5 & -0,045 & 0,269\\
-0,4 & 0,6 & -0,467 & -0,3 & 0,7 & 1 & -0,2 & -0,164 & 0,292\\
0,571 & -0,143 & 0,238 & 0,214 & -0,5 & -0,2 & 1 & 0,195 & -0,319\\
0,182 & 0,091 & 0,212 & 0,045 & -0,045 & -0,164 & 0,195 & 1 & -0,273\\
-0,154 & 0,385 & -0,333 & -0,269 & 0,269 & 0,292 & -0,319 & -0,273 & 1
\end{pmatrix}. \quad (2)$$

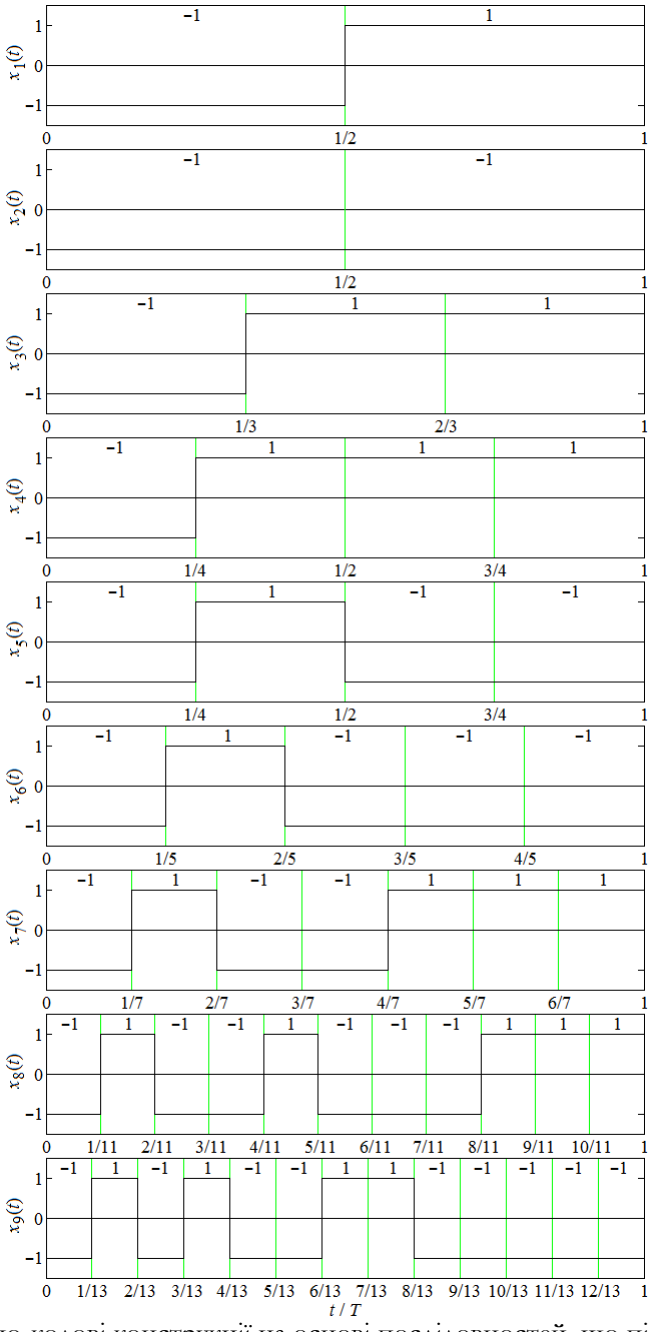

Рис. 1. Сигнально-кодові конструкції на основі послідовностей, що підлягають аналізу

Етап 3. Аналіз кореляційних зв'язків між кодовими конструкціями на основі гаусівської змішаної моделі, виділення її компонент з використанням параметрично-критеріальної модифікації ЕМ-алгоритму та структуризація виокремлених кодових конструкцій.

Для подальшого аналізу використовуватимемо лише ті елементи **R**, які знаходяться нижче її головної діагоналі, оскільки елементи, які знаходяться вище головної діагоналі, дублюють статистичну інформацію ($R_{u,v} = R_{v,u}$, $u = \overline{1,9}$, $v = \overline{1,9}$), а статистична інформація з головної діагоналі ($R_{u,u} = 1$) є тривіальною і не містить відомостей, які могли б бути корисними для подальшого аналізу кодових конструкцій. Запишемо елементи **R**, які знаходяться нижче її головної діагоналі, у спосіб, який формалізовано виразом (3).

Головною науковою гіпотезою дослідження є те, що інформація, яка міститься у матриці **R**, тобто інформація про кореляційні зв'язки між кодовими конструкціями, містить інформацію про внутрішні структури досліджуваних кодових конструкцій та взаємозв'язки між ними. Виокремити цю інформацію можна шляхом аналізу значень, які входять до **R** та, відповідно, **R** при їх представленні у вигляді змішаної гаусівської моделі [17] та подальшого аналізу й виділення (кластеризації) її компонент.

$$\mathbf{R}=\left\{R_j\right\}=\{R_{2,1},R_{3,1},R_{3,2},R_{4,1},R_{4,2},R_{4,3},R_{5,1},R_{5,2},R_{5,3},R_{5,4},R_{6,1},R_{6,2},R_{6,3},R_{6,4},R_{6,5},R_{7,1},R_{7,2},R_{7,3},$$
$$R_{7,4},R_{7,5},R_{7,6},R_{8,1},R_{8,2},R_{8,3},R_{8,4},R_{8,5},R_{8,6},R_{8,7},R_{9,1},R_{9,2},R_{9,3},R_{9,4},R_{9,5},R_{9,6},R_{9,7},R_{9,8}\}=$$
$$=\{0;\ 0,667;\ -0,333;\ 0,5;\ -0,5;\ 0,833;\ -0,5;\ 0,5;\ -0,167;$$
$$0;\ -0,4;\ 0,6;\ -0,467;\ -0,3;\ 0,7;\ 0,571;\ -0,143;\ 0,238;$$
$$0,214;\ -0,5;\ -0,2;\ 0,182;\ 0,091;\ 0,212;\ 0,045;\ -0,045;\ -0,164;$$
$$0,195;\ -0,154;\ 0,385;\ -0,333;\ -0,269;\ 0,269;\ 0,292;\ -0,319;\ -0,273\},$$
$$j=\overline{1,N},\ N=36. \quad (3)$$

Для $\mathbf{R}$ гаусівська змішана модель може бути представлена виразом (4) [17]:

$$p\left(R\right)=\sum_{k=1}^{K}\theta_k\mathcal{N}\left(R;\mu_k,\sigma_k^2\right)=\sum_{k=1}^{K}\frac{\theta_k}{\sigma_k\sqrt{2\pi}}\exp\left[-\frac{\left(R-\mu_k\right)^2}{2\sigma_k^2}\right], \quad (4)$$

де $p\left(R\right)$ – щільність розподілу ймовірностей неперервної випадкової величини R; K – кількість компонент суміші, що є гаусівською змішаною моделлю; $\mathcal{N}\left(R;\mu_k,\sigma_k^2\right)$ – k-а гаусівська компонента з математичним сподіванням μ_k та дисперсією σ_k^2, яка входить до складу суміші $p\left(R\right)$; θ_k – ваговий коефіцієнт k-ї гаусівської компоненти у складі суміші $p\left(R\right)$, який може асоціюватися з ймовірністю того, що випадкова величина R належить k-й компоненті $\mathcal{N}\left(R;\mu_k,\sigma_k^2\right)$; $\sum_{k=1}^{K}\theta_k=1$.

Зауважимо, що при $K=1$ гаусівська змішана модель перетворюється у нормальний розподіл $R\sim\mathcal{N}\left(\mu,\sigma^2\right)$ з математичним сподіванням μ та дисперсією σ^2.

При статистичному аналізі $\mathbf{R}$, що містить значення, які вказані у (3), з використанням гаусівської змішаної моделі (4) оцінюванню підлягають параметри моделі $\boldsymbol{\omega}=\{\boldsymbol{\theta},\boldsymbol{\mu},\boldsymbol{\sigma}\}=\{\theta_1,...,\theta_K,\ \mu_1,...,\mu_K,\ \sigma_1,...,\sigma_K\}$, а для розв'язання задачі кластеризації визначенню також підлягає структурний параметр моделі K, який визначатиме кількість компонент суміші (кластерів), яким належатимуть елементи $\mathbf{R}$.

Для оцінювання параметрів компонент суміші (4) часто використовується ЕМ-алгоритм (expectation-maximization algorithm) [18, 19], який за своєю сутністю реалізовує метод максимальної правдоподібності, що максимізує логарифмічну функцію правдоподібності $L\left(\boldsymbol{\omega}\,|\,\mathbf{R}\right)$ гіпотези щодо параметрів $\boldsymbol{\omega}$ при спостережуваних даних $\mathbf{R}$, які містять N об'єктів, які у задачі, що розглядається, є значеннями коефіцієнтів взаємної кореляції між досліджуваними кодовими конструкціями:

$$L\left(\boldsymbol{\omega}\,|\,\mathbf{R}\right)=\ln\prod_{j=1}^{N}p\left(R_j\,|\,\boldsymbol{\omega}\right)=$$
$$\ln\prod_{j=1}^{N}\sum_{k=1}^{K}\frac{\theta_k}{\sigma_k\sqrt{2\pi}}\exp\left[-\frac{\left(R_j-\mu_k\right)^2}{2\sigma_k^2}\right]= \quad (5)$$
$$=\sum_{j=1}^{N}\ln\sum_{k=1}^{K}\frac{\theta_k}{\sigma_k\sqrt{2\pi}}\exp\left[-\frac{\left(R_j-\mu_k\right)^2}{2\sigma_k^2}\right],$$

а оцінками параметрів компонент суміші є $\boldsymbol{\omega}=\arg\max_{\boldsymbol{\omega}}L\left(\boldsymbol{\omega}\,|\,\mathbf{R}\right)$.

ЕМ-алгоритм полягає у реалізації ітерацій з двох кроків [19, с. 252]:

1. E (expectation), на якому при поточному наближенні параметрів моделі

$$\boldsymbol{\omega}^{\langle s-1\rangle}=\left\{\boldsymbol{\theta}^{\langle s-1\rangle},\boldsymbol{\mu}^{\langle s-1\rangle},\boldsymbol{\sigma}^{\langle s-1\rangle}\right\}=$$
$$\left\{\theta_1^{\langle s-1\rangle},...,\theta_K^{\langle s-1\rangle},\mu_1^{\langle s-1\rangle},...,\mu_K^{\langle s-1\rangle},\sigma_1^{\langle s-1\rangle},...,\sigma_K^{\langle s-1\rangle}\right\},$$

де s – номер ітерації алгоритму (починаючи з $s=1$), оцінюються значення прихованих змінних

$$\gamma_{j,k}^{\langle s-1\rangle}=\frac{\frac{\theta_k^{\langle s-1\rangle}}{\sigma_k^{\langle s-1\rangle}\sqrt{2\pi}}\exp\left[-\frac{\left(R_j-\mu_k^{\langle s-1\rangle}\right)^2}{2\left(\sigma_k^{\langle s-1\rangle}\right)^2}\right]}{\sum_{l=1}^{K}\frac{\theta_l^{\langle s-1\rangle}}{\sigma_l^{\langle s-1\rangle}\sqrt{2\pi}}\exp\left[-\frac{\left(R_j-\mu_l^{\langle s-1\rangle}\right)^2}{2\left(\sigma_l^{\langle s-1\rangle}\right)^2}\right]},$$

$j=\overline{1,N}$, $k=\overline{1,K}$, які є апостеріорними ймовірностями того, що $\boldsymbol{R}_j$ належить *k*-й компоненті суміші. На цьому кроці також оцінюється кількість $N_k^{\langle s-1\rangle}$ об'єктів з $\mathbf{R}$, яка належить *k*-й компоненті суміші $N_k^{\langle s-1\rangle}=\sum_{j=1}^{N}\gamma_{j,k}^{\langle s-1\rangle}$. Вибір та обґрунтування початкового наближення параметрів моделі $\boldsymbol{\omega}^{\langle 0\rangle}=\left\{\boldsymbol{\theta}^{\langle 0\rangle},\boldsymbol{\mu}^{\langle 0\rangle},\boldsymbol{\sigma}^{\langle 0\rangle}\right\}=\left\{\theta_1^{\langle 0\rangle},...,\theta_K^{\langle 0\rangle},\mu_1^{\langle 0\rangle},...,\mu_K^{\langle 0\rangle},\sigma_1^{\langle 0\rangle},...,\sigma_K^{\langle 0\rangle}\right\}$ для випадку, який розглядається у цьому дослідженні, викладено нижче у статті.

2. M (maximization), на якому визначаються такі нові оцінки параметрів моделі

$$\boldsymbol{\omega}^{\langle s\rangle}=\left\{\boldsymbol{\theta}^{\langle s\rangle},\boldsymbol{\mu}^{\langle s\rangle},\boldsymbol{\sigma}^{\langle s\rangle}\right\}=$$

$$\left\{\theta_1^{\langle s\rangle},...,\theta_K^{\langle s\rangle},\mu_1^{\langle s\rangle},...,\mu_K^{\langle s\rangle},\sigma_1^{\langle s\rangle},...,\sigma_K^{\langle s\rangle}\right\},$$

які у поточній *s*-й ітерації максимізують $L\left(\boldsymbol{\omega}\,|\,\mathbf{R}\right)$:

$$\theta_k^{\langle s\rangle}=\frac{N_k^{\langle s-1\rangle}}{N},$$

$$\mu_k^{\langle s\rangle}=\frac{1}{N_k^{\langle s-1\rangle}}\sum_{j=1}^{N}\gamma_{j,k}^{\langle s-1\rangle}R_j,\qquad(6)$$

$$\sigma_k^{\langle s\rangle}=\sqrt{\frac{1}{N_k^{\langle s-1\rangle}}\sum_{j=1}^{N}\gamma_{j,k}^{\langle s-1\rangle}\left(R_j-\mu_k^{\langle s\rangle}\right)^2},\ k=\overline{1,K}.$$

В якості критерію зупинки ЕМ-алгоритму може бути використано умову $L\left(\boldsymbol{\omega}^{\langle s\rangle}\,|\,\mathbf{R}\right)-L\left(\boldsymbol{\omega}^{\langle s-1\rangle}\,|\,\mathbf{R}\right)<\varepsilon$, де ε – додатне число, при якому поточна точність оцінювання параметрів моделі в контексті збіжності алгоритму вважається достатньою.

В якості отриманої оцінки $\boldsymbol{\omega}$ приймається $\boldsymbol{\omega}=\boldsymbol{\omega}^{\langle s_{\max}\rangle}$.

Незважаючи на те, що ЕМ-алгоритм є в цілому відомим методом оцінювання параметрів розподілів, які утворюють суміш, та кластеризації об'єктів, одним з головних проблемних питань при його практичній реалізації є апріорна невизначеність кількості компонент (кластерів) *K*, яка є структурним параметром цього алгоритму. Відомі модифікації ЕМ-алгоритму з додаванням [20] та видаленням компонент [21]. Вибір тієї чи іншої модифікації ЕМ-алгоритму суттєвим чином пов'язаний з проблематикою задачі, яка має таку важливу особливість: кількість об'єктів у суміші $\mathbf{R}$, які підлягають аналізу та кластеризації, становить $N = 36$, що є достатньо малою кількістю для виділення декількох гаусівських компонент, навіть якщо вони присутні у спостережуваній вибірці суміші (для достатнього опису однієї, наприклад *k*-ї, гаусівської компоненти необхідно, щоб у суміші були присутні, в залежності від її параметрів μ_k та σ_k^2, принаймні $N_k>10...30$ об'єктів, які належать цій компоненті [22]). Тому використання ЕМ-алгоритму з додаванням компонент призводить до того, що значна кількість або усі (залежить від початкового наближення параметрів моделі $\boldsymbol{\omega}^{\langle 0\rangle}$) об'єкти суміші ідентифікуються з однією й тією ж самою компонентою (одна або дві початкові компоненти при кластеризації є "аттракторами" усіх об'єктів суміші при малому об'ємі їх вибірки *N*), а процес додавання компонент є суттєво ускладненим через необхідність більш складного обґрунтування критерію додавання компонент на основі, наприклад, правила $\forall k\in\overline{1,K}$, $\gamma_{j,k}<\gamma_0$, де γ_0 – порогове значення апостеріорної ймовірності того, що $\boldsymbol{R}_j$ належить якійсь компоненті з поточних існуючих *K* компонент суміші (тобто існує проблема вибору порогу γ_0, від якого, у свою чергу, залежить ймовірність помилок першого та другого роду при кластеризації).

Тому для розв'язання задачі дослідження цієї статті було обрано модифікацію ЕМ-алгоритму з видаленням компонент. Ця модифікація ЕМ-алгоритму в якості структурного параметру має початкову максимально можливу кількість компонент $K_{\max}$.

Параметрично-критеріальна модифікація ЕМ-алгоритму з видаленням компонент для розв'язання задачі дослідження: обґрунтування структурного параметру $K_{\max}$, початкових наближень параметрів моделі $\boldsymbol{\omega}^{\langle 0\rangle}$ та додаткових критеріїв кластеризації ЕМ-алгоритму з видаленням компонент.

Враховуючи те, що аналізу підлягають дані $\mathbf{R}$, які є $N = 36$ значеннями коефіцієнтів взаємної кореляції між кодовими конструкціями, можна було б обрати максимально можливу кількість компонент суміші $K_{\max} = N = 36$. Однак розв'язання задачі кластеризації у цьому випадку в залежності від початкового наближення параметрів моделі $\boldsymbol{\omega}^{\langle 0\rangle}$ може зводитися до того, що є 36 кластерів, у кожному з яких знаходиться по одному або декілька (якщо їх значення співпадають, як у випадку, наприклад, $R_1 = R_{10}$) коефіцієнтів взаємної кореляції

$\boldsymbol{R}_j$, $j=\overline{1,N}$. При цьому буде також певна кількість порожніх кластерів, оскільки $K_{\max} = N$, а до якось кластеру буде входити більше одного об'єкта. Зміст та інтерпретація такої кластеризації є тривіальними і показують лише те, що будь-які дві різні кодові конструкції певним чином корельовані між собою, що є апріорі відомим, або декілька різних кодових конструкцій корельовані між собою попарно однаково з одним і тим самим значенням коефіцієнта кореляції, що також можна визначити під час простого аналізу значень $\mathbf{R}$. Для того щоб проаналізувати структуру певної кодової конструкції в контексті її лінійних зв'язків з іншими кодовими конструкціями досліджуваної системи, необхідно висунути та дослідити гіпотезу про те, що до кластеру входить принаймні два об'єкти, тому в якості початкового структурного параметру ЕМ-алгоритму з видаленням компонент оберемо $K_{\max} = N/2 = 18$.

Початкове наближення доцільно обрати таким чином щоб забезпечити максимально можливу еквідистантність між гаусівськими компонентами у діапазоні усіх можливих значень коефіцієнта взаємної кореляції у загальному випадку, тобто $-1 \le R \le 1$. Це забезпечить рівномірне охоплення початковими кластерами діапазону усіх можливих значень R під час аналізу. Положення початкових кластерів при цьому визначатиметься $\boldsymbol{\mu}^{\langle 0\rangle}$, які легко визначити з умови максимально можливої еквідистантності між K кластерами:

$$\mu_k^{\langle 0\rangle} = -1 + k\Delta\mu - \frac{\Delta\mu}{2},\ k=\overline{1,K},$$

де $\Delta\mu = \dfrac{\max R - \min R}{K} = \dfrac{2}{K}$.

Ширину одного початкового кластеру можна визначити, використовуючи правило 3σ, при якому кожен початковий кластер матиме інтервал $\left(\mu_k^{\langle 0\rangle} - 3\sigma_k^{\langle 0\rangle}; \mu_k^{\langle 0\rangle} + 3\sigma_k^{\langle 0\rangle}\right)$. Враховуючи умову максимально можливої еквідистантності між K кластерами та правило 3σ для ширини одного початкового кластеру, значення $\sigma_k^{\langle 0\rangle}$ визначатиметься таким чином:

$$\sigma_k^{\langle 0\rangle} = \frac{\max R - \min R}{6K} = \frac{1}{3K},\ k=\overline{1,K}.$$

В умовах апріорної невизначеності щодо розподілу об'єктів по початковим кластерам вважатимемо що ці об'єкти розподіляються між K кластерами рівномірно, тобто

$$\theta_k^{\langle 0\rangle} = \frac{1}{K},\ k=\overline{1,K}.$$

Реалізація ЕМ-алгоритму фактично полягає в уточненні зазначених початкових наближень параметрів моделі $\boldsymbol{\omega}^{\langle 0\rangle} = \left\{\boldsymbol{\theta}^{\langle 0\rangle}, \boldsymbol{\mu}^{\langle 0\rangle}, \boldsymbol{\sigma}^{\langle 0\rangle}\right\}$ таким чином, що на кожній новій s-й його ітерації отримуватимуться такі нові уточнені наближення $\boldsymbol{\omega}^{\langle s\rangle} = \left\{\boldsymbol{\theta}^{\langle s\rangle}, \boldsymbol{\mu}^{\langle s\rangle}, \boldsymbol{\sigma}^{\langle s\rangle}\right\}$, які є більш правдоподібними в умовах спостережуваних даних $\mathbf{R}$, ніж наближення, отримані у попередніх ітераціях (функція правдоподібності $L\left(\boldsymbol{\omega} \mid \mathbf{R}\right)$ зростатиме у кожній ітерації).

Введемо такі три додаткові критерії кластеризації для ЕМ-алгоритму з видаленням компонент, які дозволяють приймати рішення про корегування його подальшої реалізації при досягненні певних умов. Запропоновані критерії враховують особливості контексту кластеризації об'єктів, яка здійснюється з метою подальшої структуризації кодових конструкцій.

<u>Критерій 1</u>. *k-й кластер* (гаусівська компонента) *є порожнім (не містить об'єктів) та повинен бути видалений при подальшому аналізі* тоді, коли на s-у кроці ЕМ-алгоритму спостерігається одночасне виконання таких умов, які є взаємопов'язаними через апостеріорні ймовірності $\gamma_{j,k}^{\langle s-1\rangle}$ ($\gamma_{j,k}^{\langle s-1\rangle} \to 0$, $j=\overline{1,N}$, при виконанні нижчезазначених жорстких умов):

1) $N_k^{\langle s-1\rangle} \ll 1$ (жорстка умова: $N_k^{\langle s-1\rangle} = 0$) – визначальна умова;

2) $\sigma_k^{\langle s\rangle} \approx 0$ (жорстка умова: $\sigma_k^{\langle s\rangle} = 0$);

3) $\theta_k^{\langle s\rangle} \ll \dfrac{1}{K}$ (жорстка умова: $\theta_k^{\langle s\rangle} = 0$).

Зазначимо, що при практичній реалізації ЕМ-алгоритму вказані вище жорсткі умови критерію 1 можуть не досягатися точно через те, що оцінювані апостеріорні ймовірності $\gamma_{j,k}^{\langle s-1\rangle}$, $j=\overline{1,N}$, для k-го кластеру, який не містить елементів, не завжди становлять точні значення $\gamma_{j,k}^{\langle s-1\rangle} = 0$, $j=\overline{1,N}$, а можуть бути наближеними до них з прийнятною для прийняття рішення відповідно до критерію 1 точністю оцінками, тобто $\gamma_{j,k}^{\langle s-1\rangle} \approx 0$, $j=\overline{1,N}$.

<u>Критерій 2</u>. *k-й кластер* (гаусівська компонента) *містить лише один об'єкт та повинен бути видалений при подальшому аналізі* через те, що цей один виділений об'єкт (коефіцієнт взаємної кореляції) у кластері є тривіальним (апріорі відомо, що будь-

які дві різні кодові конструкції певним чином корельовані між собою; див. також обґрунтування структурного параметру $K_{\max}$ вище) тоді, коли на s-у кроці ЕМ-алгоритму спостерігається одночасне виконання таких умов:

1) $N_k^{\langle s-1\rangle} \approx 1$ (жорстка умова: $N_k^{\langle s-1\rangle} = 1$) – визначальна умова;

2) $\sigma_k^{\langle s\rangle} \approx 0$ (жорстка умова: $\sigma_k^{\langle s\rangle} = 0$).

Об'єкт, який виділено відповідно до критерію 2, залишається у складі $\mathbf{R}$ *для подальшого аналізу* (подальший аналіз відбувається без видаленого k-го кластеру).

Критерій 3. *k-й кластер* (гаусівська компонента) *містить декілька однакових об'єктів*:

1) $N_k^{\langle s-1\rangle} > 1$ (жорстка умова: $N_k^{\langle s-1\rangle} > 1$, $N_k^{\langle s-1\rangle} \in \mathbb{Z}_+$) – визначальна умова;

2) $\sigma_k^{\langle s\rangle} \approx 0$ (жорстка умова: $\sigma_k^{\langle s\rangle} = 0$).

Кодові конструкції, що беруть участь у формуванні об'єктів, які виділено за критерієм 3, вилучаються та аналізуються на предмет їх можливої структуризації. *Подальший аналіз проводиться для даних* $\mathbf{R}$, *які формуються без вилучених кодових конструкцій.*

Введення для розв'язання поставленої задачі вказаних вище критеріїв пов'язане з тим, що при вказаних для них жорстких умов виконується $\sigma_k^{\langle s\rangle} = 0$ для певного k-го кластеру, що вводить ЕМ-алгоритм в стан математичної сингулярності з утворенням невизначеностей типу $L\left(\boldsymbol{\omega}^{\langle s\rangle} \mid \mathbf{R}\right) = 0/0$, що випливає з аналізу (5). Введені критерії дозволяють виходити з зазначеного стану математичної сингулярності та продовжувати подальший статистичний аналіз з кластеризацію об'єктів. При цьому введені критерії також пояснюють значення умов, якими супроводжується вхід ЕМ-алгоритму в стан математичної сингулярності в контексті розв'язуваної задачі.

У цій статті не наводимо повний математичний аналіз математичних сингулярностей ЕМ-алгоритму з розкриттям відповідних невизначеностей типу $L\left(\boldsymbol{\omega}^{\langle s\rangle} \mid \mathbf{R}\right) = 0/0$, проте зазначимо, що:

1) при виконанні умов критерію 1 ця невизначеність розкриватиметься таким чином, що значення $L\left(\boldsymbol{\omega}^{\langle s\rangle} \mid \mathbf{R}\right)$ прямуватиме до такого, яке було б без порожнього k-го кластеру, що слідує з (5) при відповідному ваговому коефіцієнті k-ї гаусівської компоненти $\theta_k = 0$;

2) при виконанні умов критерію 2 або критерію 3 $L\left(\boldsymbol{\omega}^{\langle s\rangle} \mid \mathbf{R}\right) \to +\infty$, що можна пояснити так: при виконанні критерію 2 або критерію 3 у складі суміші ідентифікується k-а вироджена гаусівська компонента $\mathcal{N}\left(R; \mu_k, 0\right)$, яка містить у своєму складі лише декілька (або одне у випадку критерію 2) однакових значень (констант); ці константи дорівнюють їх середньому μ_k, а класична дзвоноподібна функція щільності розподілу ймовірностей $p\left(R\right) = \dfrac{1}{\sigma_k\sqrt{2\pi}} \exp\left[-\dfrac{\left(R-\mu_k\right)^2}{2\sigma_k^2}\right]$ вироджується в цьому випадку у δ-функцію Дірака $p\left(R \mid \sigma_k \to 0\right) = \delta\left(R-\mu_k\right)$; якщо така необмежена зверху $p\left(R\right)$ входить до складу $L\left(\boldsymbol{\omega}^{\langle s\rangle} \mid \mathbf{R}\right)$ як вироджена гаусівська компонента, то підстановка у $L\left(\boldsymbol{\omega}^{\langle s\rangle} \mid \mathbf{R}\right)$ спостережуваного значення $R_j = \mu_k$ дає $L\left(\boldsymbol{\omega}^{\langle s\rangle} \mid \mathbf{R}\right) \to +\infty$.

Реалізація параметрично-критеріальної модифікації ЕМ-алгоритму з видаленням компонент для розв'язання задачі структуризації кодових конструкцій.

Раунд 1.

Вхідні дані: $\mathbf{R}$, $N = 36$.

Структурний параметр $K = N/2 = 18$.

Початкове наближення $\boldsymbol{\omega}^{\langle 0\rangle}$:

$$\theta_k^{\langle 0\rangle} = \frac{1}{K} = \frac{1}{18},\ \mu_k^{\langle 0\rangle} = -1 + \frac{2k-1}{K} = -1 + \frac{2k-1}{18},$$

$$\sigma_k^{\langle 0\rangle} = \frac{1}{3K} = \frac{1}{54},\ k = \overline{1,18}.$$

Процес поітераційної максимізації $L\left(\boldsymbol{\omega} \mid \mathbf{R}\right)$: $L\left(\boldsymbol{\omega}^{\langle 0\rangle} \mid \mathbf{R}\right) = -36{,}1$, $L\left(\boldsymbol{\omega}^{\langle 1\rangle} \mid \mathbf{R}\right) = 28{,}1$, $L\left(\boldsymbol{\omega}^{\langle 2\rangle} \mid \mathbf{R}\right) = 0/0$ (досягнення математичної сингулярності ЕМ-алгоритму через виконання умов критеріїв 1 – 3 з розкриттям невизначеності при прийнятті до уваги "жорстких умов" критерію 2 та критерію 3 $L\left(\boldsymbol{\omega}^{\langle 2\rangle} \mid \mathbf{R}\right) \to +\infty$).

Отримані результати:

1) оцінки параметрів гаусівської змішаної моделі $\boldsymbol{\omega}^{\langle 2\rangle}$:

$$\boldsymbol{\theta}^{\langle 2\rangle} = \{0;\ 0;\ 0;\ 2\cdot 10^{-7};\ 0{,}11;\ 0{,}05;\ 0{,}14;\ 0{,}14;\ 0{,}06;\ 0{,}08;\ 0{,}11;\ 0{,}09;\ 0{,}03;\ 0{,}06;\ 0{,}06;\ 0{,}05;\ 7\cdot 10^{-4};\ 0{,}03\};$$

$$\boldsymbol{\mu}^{\langle 2\rangle} = \{-0{,}5;\ -0{,}5;\ -0{,}5;\ -0{,}5;\ -0{,}492;\ -0{,}368;\ -0{,}301;\ -0{,}165;\ -0{,}02;$$
$$0{,}046;\ 0{,}201;\ 0{,}265;\ 0{,}385;\ 0{,}5;\ 0{,}596;\ 0{,}686;\ 0{,}833;\ 0{,}833\};$$

$$\boldsymbol{\sigma}^{\langle 2\rangle} = \{0;\ 0;\ 0;\ 0;\ 0{,}014;\ 0{,}037;\ 0{,}026;\ 0{,}019;\ 0{,}023;\ 0{,}039;\ 0{,}014;\ 0{,}024;\ 0;\ 0;\ 0{,}031;\ 0{,}016;\ 0;\ 0\};$$

2) оцінка кількості об'єктів у кластерах:

$$\mathbf{N}^{\langle 1\rangle} = \{0;\ 0;\ 0;\ 7\cdot 10^{-6};\ 3{,}99;\ 1{,}85;\ 5{,}17;\ 5;\ 2{,}11;\ 2{,}89;\ 3{,}94;\ 3{,}06;\ 1;\ 2;\ 2{,}29;\ 1{,}72;\ 0{,}03;\ 0{,}97\}.$$

Аналіз отриманих у раунді 1 результатів:

1) Критерію 1 відповідають 5 кластерів: $k = 1$; $k = 2$; $k = 3$; $k = 4$; $k = 17$. Ці порожні кластери видаляються з подальшого аналізу.

2) Критерію 2 відповідають два кластери: $k = 13$; $k = 18$. У кластері $k = 13$ виокремлено об'єкт R_{30} $(\gamma^{\langle 1\rangle}_{30,13} = 1)$, у кластері $k = 18$ виокремлено об'єкт R_6 $(\gamma^{\langle 1\rangle}_{6,18} = 0{,}973)$. Ці кластери, які містять по одному елементу, видаляються з подальшого аналізу. Об'єкти, які було у них виокремлено, залишаються для подальшого аналізу.

3) Критерію 3 відповідає один кластер $k = 14$, у якому виокремлено об'єкти R_4 $(\gamma^{\langle 1\rangle}_{4,14} = 0{,}998)$ та R_8 $(\gamma^{\langle 1\rangle}_{8,14} = 0{,}998)$. Цей кластер, який містить два об'єкти, видаляється з подальшого аналізу. Об'єкти, які було у ньому виокремлено, а також відповідні ним кодові конструкції, вилучаються з подальшого аналізу у ЕМ-алгоритмі та підлягають аналізу на предмет можливої структуризації.

Зазначене вище видалення кластерів може бути здійснено шляхом встановлення нових наближень (перепризначення кластерів) у наступному раунді ЕМ-алгоритму.

Аналіз виокремлених об'єктів: $R_4 = R_{4,1} = 0{,}5$, $R_8 = R_{5,2} = 0{,}5$.

Сигнально-кодові конструкції, які відповідають виокремленим об'єктам:

$\mathbf{X}_1$ та $\mathbf{X}_4$ – для об'єкта $R_4 = R_{4,1} = 0{,}5$;

$\mathbf{X}_2$ та $\mathbf{X}_5$ – для об'єкта $R_8 = R_{5,2} = 0{,}5$.

Виявлена можлива структуризація виокремлених у першому раунді кластеризації з використанням ЕМ-алгоритму кодових конструкцій $\{\mathbf{X}_1, \mathbf{X}_4\}$ та $\{\mathbf{X}_2, \mathbf{X}_5\}$ (довжини послідовностей $D = 2$ та $D = 4$) показана на рис. 2.

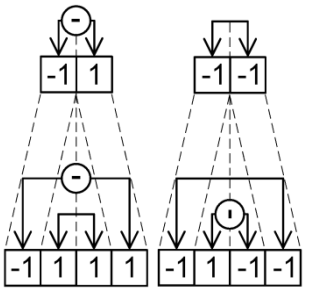

Рис. 2. Результат структуризації кодових конструкцій при $D = 2$ та $D = 4$ у першому раунді ЕМ-алгоритму

Раунд 2.

Вхідні дані. З урахуванням того, що на першому раунді ЕМ-алгоритму були виокремлені та структуризовані кодові конструкції $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_4$, та $\mathbf{X}_5$, подальшому аналізу підлягає система кодових конструкцій $\mathbf{X}_3$, $\mathbf{X}_6$, $\mathbf{X}_7$, $\mathbf{X}_8$, та $\mathbf{X}_9$.

Матриця коефіцієнтів взаємної кореляції для цієї системи кодових конструкцій утворюється з матриці $\mathbf{R}$ шляхом видалення строк та стовпчиків з номерами 1, 2, 4, 5. Утворена таким чином матриця коефіцієнтів взаємної кореляції $\mathbf{R}^*$ представлена у (7):

$$\mathbf{R}^* = \begin{pmatrix} 1 & -0{,}467 & 0{,}238 & 0{,}212 & -0{,}333 \\ -0{,}467 & 1 & -0{,}2 & -0{,}164 & 0{,}292 \\ 0{,}238 & -0{,}2 & 1 & 0{,}195 & -0{,}319 \\ 0{,}212 & -0{,}164 & 0{,}195 & 1 & -0{,}273 \\ -0{,}333 & 0{,}292 & -0{,}319 & -0{,}273 & 1 \end{pmatrix}, \qquad (7)$$

$$\mathbf{R}^* = \left\{R_j^*\right\} = \left\{R_{2,1}^*,\ R_{3,1}^*,\ R_{3,2}^*,\ R_{4,1}^*,\ R_{4,2}^*,\ R_{4,3}^*,\ R_{5,1}^*,\ R_{5,2}^*,\ R_{5,3}^*,\ R_{5,4}^*\right\} =$$

$$= \{-0,467;\ 0,238;\ -0,2;\ 0,212;\ -0,164;\ 0,195;\ -0,333;\ 0,292;\ -0,319;\ -0,273\},$$

$$j = \overline{1,N},\ N = 10.$$

Структурний параметр $K = N/2 = 5$.

Початкове наближення $\boldsymbol{\omega}^{\langle 0\rangle}$: $\theta_k^{\langle 0\rangle} = \frac{1}{K} = \frac{1}{5}$, $\mu_k^{\langle 0\rangle} = -1 + \frac{2k-1}{K} = -1 + \frac{2k-1}{5}$, $\sigma_k^{\langle 0\rangle} = \frac{1}{3K} = \frac{1}{15}$, $k = \overline{1,5}$.

Процес поітераційної максимізації $L\left(\boldsymbol{\omega} \mid \mathbf{R}^*\right)$:

$L\left(\boldsymbol{\omega}^{\langle 0\rangle} \mid \mathbf{R}^*\right) = -20,2$, $L\left(\boldsymbol{\omega}^{\langle 1\rangle} \mid \mathbf{R}^*\right) = 4,5$, $L\left(\boldsymbol{\omega}^{\langle 2\rangle} \mid \mathbf{R}^*\right) = 0/0$ (досягнення математичної сингулярності ЕМ-алгоритму через виконання умов критерію 1 з розкриттям невизначеності при прийнятті до уваги його "жорстких умов" $L\left(\boldsymbol{\omega}^{\langle 2\rangle} \mid \mathbf{R}^*\right) = 5,5165$).

Отримані результати:

1) оцінки параметрів гаусівської змішаної моделі $\boldsymbol{\omega}^{\langle 2\rangle}$:

$$\boldsymbol{\theta}^{\langle 2\rangle} = \left\{1,1\cdot 10^{-4};\ 0,479;\ 0,156;\ 0,365;\ 5\cdot 10^{-13}\right\};$$

$$\boldsymbol{\mu}^{\langle 2\rangle} = \left\{-0,467;\ -0,312;\ -0,115;\ 0,235;\ 0,292\right\};$$

$$\boldsymbol{\sigma}^{\langle 2\rangle} = \left\{0;\ 0,094;\ 0,196;\ 0,037;\ 1,5\cdot 10^{-14}\right\};$$

2) оцінка кількості об'єктів у кластерах:

$$\mathbf{N}^{\langle 1\rangle} = \left\{1,1\cdot 10^{-3};\ 4,789;\ 1,562;\ 3,648;\ 5\cdot 10^{-12}\right\}.$$

Аналіз отриманих у раунді 2 результатів:

1) Критерію 1 відповідають 2 кластери: $k = 1$ та $k = 5$. Ці порожні кластери видаляються з подальшого аналізу.

2) Критеріям 2 та 3 не відповідає жоден з кластерів.

Раунд 3.

Вхідні дані: $\mathbf{R}^*$, $N = 10$.

Структурний параметр $K = 3$ (через видалення двох порожніх кластерів у попередньому раунді).

Початкове наближення $\boldsymbol{\omega}^{\langle 0\rangle}$ (відповідає останньому отриманому наближенню $\boldsymbol{\omega}^{\langle 2\rangle}$ у попередньому раунді, але без порожніх кластерів):

$$\boldsymbol{\theta}^{\langle 0\rangle} = \{0,479;\ 0,156;\ 0,365\};$$

$$\boldsymbol{\mu}^{\langle 0\rangle} = \{-0,312;\ -0,115;\ 0,235\};$$

$$\boldsymbol{\sigma}^{\langle 0\rangle} = \{0,094;\ 0,196;\ 0,037\}.$$

Процес поітераційної максимізації $L\left(\boldsymbol{\omega} \mid \mathbf{R}^*\right)$:

$L\left(\boldsymbol{\omega}^{\langle 0\rangle} \mid \mathbf{R}^*\right) = 5,5165$, $L\left(\boldsymbol{\omega}^{\langle 1\rangle} \mid \mathbf{R}^*\right) = 5,9178$,

$L\left(\boldsymbol{\omega}^{\langle 2\rangle} \mid \mathbf{R}^*\right) = 6,1345$, $L\left(\boldsymbol{\omega}^{\langle 3\rangle} \mid \mathbf{R}^*\right) = 6,1836$, …,

$L\left(\boldsymbol{\omega}^{\langle 30\rangle} \mid \mathbf{R}^*\right) = 7,1722$, $L\left(\boldsymbol{\omega}^{\langle 31\rangle} \mid \mathbf{R}^*\right) = 7,1724$,

$L\left(\boldsymbol{\omega}^{\langle 32\rangle} \mid \mathbf{R}^*\right) = 7,1725$, $L\left(\boldsymbol{\omega}^{\langle 33\rangle} \mid \mathbf{R}^*\right) = 7,1725$

(зупинка ЕМ-алгоритму при виконанні умови $L\left(\boldsymbol{\omega}^{\langle s\rangle} \mid \mathbf{R}^*\right) - L\left(\boldsymbol{\omega}^{\langle s-1\rangle} \mid \mathbf{R}^*\right) < \varepsilon$ при $\varepsilon = 10^{-4}$).

Отримані результати:

1) оцінки параметрів гаусівської змішаної моделі $\boldsymbol{\omega}^{\langle 33\rangle}$:

$$\boldsymbol{\theta}^{\langle 33\rangle} = \{0,425;\ 0,175;\ 0,4\};$$

$$\boldsymbol{\mu}^{\langle 33\rangle} = \{-0,339;\ -0,181;\ 0,234\};$$

$$\boldsymbol{\sigma}^{\langle 33\rangle} = \{0,079;\ 0,018;\ 0,037\};$$

2) оцінка кількості об'єктів у кластерах:

$$\mathbf{N}^{\langle 32\rangle} = \{4,246;\ 1,754;\ 4\}.$$

Аналіз отриманих у раунді 3 результатів:

1) Критеріям 1 – 3 не відповідає жоден з кластерів.

2) Аналізу підлягають об'єкти, які виділені у трьох кластерах раунду 3:

– у кластері $k = 1$ виокремлено 4 об'єкти: R_1^* ($\gamma_{1,1}^{\langle 32\rangle} = 1$), R_7^* ($\gamma_{7,1}^{\langle 32\rangle} = 1$), R_9^* ($\gamma_{9,1}^{\langle 32\rangle} = 1$), R_{10}^* ($\gamma_{10,1}^{\langle 32\rangle} = 1$);

– у кластері $k = 2$ виокремлено 2 об'єкти: R_3^* ($\gamma_{3,2}^{\langle 32\rangle} = 0,825$), R_5^* ($\gamma_{5,2}^{\langle 32\rangle} = 0,929$);

– у кластері $k = 3$ виокремлено 4 об'єкти: R_2^* ($\gamma_{2,3}^{\langle 32\rangle} = 1$), R_4^* ($\gamma_{4,3}^{\langle 32\rangle} = 1$), R_6^* ($\gamma_{6,3}^{\langle 32\rangle} = 1$), R_8^* ($\gamma_{8,3}^{\langle 32\rangle} = 1$).

Аналіз виокремлених об'єктів:

– у кластері $k = 1$: $R_1^* = R_{2,1}^* = R_{6,3} = -0,467$, $R_7^* = R_{5,1}^* = R_{9,3} = -0,333$, $R_9^* = R_{5,3}^* = R_{9,7} = -0,319$, $R_{10}^* = R_{5,4}^* = R_{9,8} = -0,273$;

– у кластері $k = 2$: $R_3^* = R_{3,2}^* = R_{7,6} = -0,2$, $R_5^* = R_{4,2}^* = R_{8,6} = -0,164$;

– у кластері $k = 3$: $R_2^* = R_{3,1}^* = R_{7,3} = 0,238$, $R_4^* = R_{4,1}^* = R_{8,3} = 0,212$, $R_6^* = R_{4,3}^* = R_{8,7} = 0,195$, $R_8^* = R_{5,2}^* = R_{9,6} = 0,292$.

Кодові конструкції, які відповідають виокремленим об'єктам у кластері $k = 1$ раунду 3 та результат їх можливої структуризації:

$\mathbf{X}_3$ та $\mathbf{X}_6$ – для об'єкта $R_1^* = R_{2,1}^* = R_{6,3} = -0,467$;

$\mathbf{X}_3$ та $\mathbf{X}_9$ – для об'єкта $R_7^* = R_{5,1}^* = R_{9,3} = -0,333$;

$\mathbf{X}_7$ та $\mathbf{X}_9$ – для об'єкта $R_9^* = R_{5,3}^* = R_{9,7} = -0,319$;

$\mathbf{X}_8$ та $\mathbf{X}_9$ – для об'єкта $R_{10}^* = R_{5,4}^* = R_{9,8} = -0,273$.

Можлива структуризація кодових конструкцій у кластері $k = 1$ раунду 3 не виявлена.

Кодові конструкції, які відповідають виокремленим об'єктам у кластері $k = 2$ раунду 3 та результат їх можливої структуризації:

$\mathbf{X}_6$ та $\mathbf{X}_7$ – для об'єкта $R_3^* = R_{3,2}^* = R_{7,6} = -0,2$;

$\mathbf{X}_6$ та $\mathbf{X}_8$ – для об'єкта $R_5^* = R_{4,2}^* = R_{8,6} = -0,164$.

Можлива структуризація кодових конструкцій у кластері $k = 2$ раунду 3 не виявлена.

Кодові конструкції, які відповідають виокремленим об'єктам у кластері $k = 3$ раунду 3 та результат їх можливої структуризації:

$\mathbf{X}_3$ та $\mathbf{X}_7$ – для об'єкта $R_2^* = R_{3,1}^* = R_{7,3} = 0,238$;

$\mathbf{X}_3$ та $\mathbf{X}_8$ – для об'єкта $R_4^* = R_{4,1}^* = R_{8,3} = 0,212$;

$\mathbf{X}_7$ та $\mathbf{X}_8$ – для об'єкта $R_6^* = R_{4,3}^* = R_{8,7} = 0,195$;

$\mathbf{X}_6$ та $\mathbf{X}_9$ – для об'єкта $R_8^* = R_{5,2}^* = R_{9,6} = 0,292$.

Виявлена можлива структуризація виокремлених у третьому раунді кластеризації з використанням ЕМ-алгоритму кодових конструкцій $\{\mathbf{X}_3, \mathbf{X}_7, \mathbf{X}_8\}$ та $\{\mathbf{X}_6, \mathbf{X}_9\}$ (довжини послідовностей $D = 3$, $D = 7$, $D = 11$ та $D = 5$, $D = 13$) показана на рис. 3 та рис. 4.

Зауважимо, що у раунді 3 ЕМ-алгоритму у кластерах $k = 1$ та $k = 2$ було виокремлено побічні результати статистичного аналізу, які не дали результату щодо можливої структуризації кодових конструкцій. Ознакою того, що об'єкти, які містять інформацію щодо структуризації кодових конструкцій, знаходяться саме у кластері $k = 3$ є той факт, що цей кластер було виокремлено ЕМ-алгоритмом у раунді 3 найбільш явно: $N_3^{\langle 32 \rangle} = 4$ (у порівнянні з $N_1^{\langle 32 \rangle} = 4,246$ та $N_2^{\langle 32 \rangle} = 1,754$ для кластеру $k = 1$ та $k = 2$ відповідно) за умови отриманих оцінок апостеріорних ймовірностей приналежності йому кластеризованих об'єктів $\gamma_{2,3}^{\langle 32 \rangle} = \gamma_{4,3}^{\langle 32 \rangle} = \gamma_{6,3}^{\langle 32 \rangle} = \gamma_{8,3}^{\langle 32 \rangle} = 1$. Ще однією важливою ознакою кластеризації у кластері $k = 3$ є те, що в ньому нескладно ідентифікувати субкластер з об'єктів, якими є усі можливі попарні взаємні кореляції у підсистемі кодових конструкцій $\{\mathbf{X}_3, \mathbf{X}_7, \mathbf{X}_8\}$.

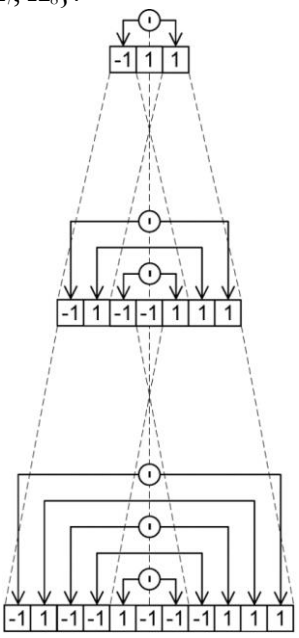

Рис. 3. Результат структуризації кодових конструкцій при $D = 3$, $D = 7$ та $D = 11$ у третьому раунді ЕМ-алгоритму

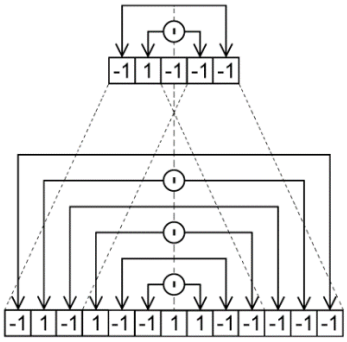

Рис. 4. Результат структуризації кодових конструкцій при $D = 5$ та $D = 13$ у третьому раунді ЕМ-алгоритму

Залежно від постановки задачі структуризації в якості вхідних даних для аналізу у запропонованому методі можуть також використовуватися більш складні досліджувані об'єкти, наприклад ПВП у широкосмугових системах DSSS [23], мультиплікативно комплементарні бінарні сигнально-кодові конструкції [24], структури досліджуваних кореляційних функцій ПВП [25] тощо.

Висновки

У статті запропоновано та досліджено особливості реалізації нового методу структуризації кодових конструкцій з апріорі невідомою для сторони здійснення аналізу структурою. Запропонований метод використовує для поглибленого статистичного аналізу гаусівську змішану модель щодо кореляційних зв'язків між досліджуваними кодовими конструкціями та параметрично-критеріальну модифікацію ЕМ-алгоритму з видаленням компонент для виділення тих компонент суміші у вигляді гаусівської змішаної моделі, які в контексті розв'язуваної задачі мають зміст інформативно значущих для структуризації досліджуваних кодових конструкцій кластерів, тобто містять інформацію щодо внутрішніх структур кодових конструкцій та взаємозв'язків між ними.

На прикладі розв'язання задачі структуризації відомих бінарних послідовностей Баркера, які можна вважати кодовими конструкціями з апріорі стохастичним неструктурованим характером, показано реалізацію запропонованого методу та отримано результати структуризації (виявлені структури кодових конструкцій та взаємозв'язки між ними).

До важливих особливостей запропонованого методу можна віднести такі:

– метод чутливий до різних інверсно-ізоморфних структур досліджуваних кодових конструкцій, що визначає його певний емпіричний та комбінаторний характер на підготовчому етапі попереднього аналізу та підготовки вхідних даних (систем досліджуваних кодових конструкцій) стороною здійснення аналізу;

– при практичній програмній реалізації методу можуть виникати математичні сингулярності у використовуваній у ньому модифікації ЕМ-алгоритму, які легко ідентифікуються повідомленнями типу "ділення на 0" (MATLAB або інші пакети моделювання); вихід з цих сингулярностей та їх зміст в контексті розв'язуваної задачі забезпечуються та здійснюються з використанням трьох введених та описаних у методі додаткових критеріїв кластеризації.

Запропонований метод може бути використано як інструмент додаткового поглибленого статистичного аналізу різних кодових конструкцій (S-блоків, ПВП тощо) на предмет виявлення їх структурованості у криптографічних та стеганографічних методах захисту інформації з метою подальшого аналізу впливу виявлених структурованостей і взаємозв'язків між ними на рівень стійкості до атак, зокрема до лінійного криптоаналізу, цих методів захисту інформації.

ЛІТЕРАТУРА

[1]. С. Остапов, С. Євсеєв, О. Король, *Технології захисту інформації: навч. посіб.*, Харків: Вид. ХНЕУ, 2013, 476 с.

[2]. В. Ємець, А. Мельник, Р. Попович, *Сучасна криптографія. Основні поняття*, Львів: БаК, 2003, 144 с.

[3]. А. Алексейчук, Л. Ковальчук, А. Шевцов, С. Яковлев, "О криптографических свойствах нового национального стандарта шифрования Украины", *Кибернетика и системный анализ*, Т. 52, № 3, С. 16-31, 2016.

[4]. ДСТУ 7624:2014. Інформаційні технології. Криптографічний захист інформації. Алгоритм симетричного блокового перетворення. – Введ. 01–07–2015. – К.: Мінекономрозвитку України, 2015.

[5]. Р. Олійников, І. Горбенко, О. Казимиров, В. Руженцев, Ю. Горбенко, "Принципи побудови і основні властивості нового національного стандарту блокового шифрування України", *Захист інформації*, Т. 17, № 2, С. 142-157, 2015.

[6]. И. Горбенко, В. Долгов, И. Лисицкая, Р. Олейников, "Новая идеология оценки стойкости блочных симметричных шифров к атакам дифференциального и линейного криптоанализа", *Прикладная радиоэлектроника*, Т. 9, № 3, С. 312-320, 2010.

[7]. Л. Бабенко, Е. Ищукова, "Особенности применения методов линейного и дифференциального криптоанализа к симметричным блочным шифрам", *Вопросы кибербезопасности*, № 1(9), С. 11-19, 2015.

[8]. Е. Фауре, С. Сисоєнко, "Метод підвищення стійкості псевдовипадкових послідовностей до лінійного криптоаналізу", *The scientific potential of the present*, Dec. 1, 2016 (St. Andrews, Scotland, UK), Proceedings, С. 119-122, 2016.

[9]. С. Евсеев, С. Остапов, И. Белодед, "Исследования свойств гибридных крипто-кодовых конструкций", *Захист інформації*, Т. 19, № 4, С. 278-290, 2017.

[10]. А. Смирнов, *Методы и средства компьютерной стеганографии с применением сложных дискретных сигналов для защиты информации в компьютерных сетях: монография*, Кировоград: "КОД", 2012, 352 с.

[11]. NIST SP 800-22 Rev. 1a. A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications. Apr. 2010. DOI: 10.6028/NIST.SP.800-22r1a

[12]. М. Мазурков, *Основи теорії передавання інформації: навч. посіб. для студ. вищ. навч. закл.*, Одеса: Наука і техніка, 2005, 168 с.

[13]. В. Бабак, А. Білецький, *Детерміновані сигнали і спектри: навч. посіб. для студ. вищ. навч. закл.*, К.: Техніка, 2003, 455 с.

[14]. В. Гантмахер, Н. Быстров, Д. Чеботарев, *Шумоподобные сигналы. Анализ, синтез, обработка*, СПб.: Наука и техника, 2005, 400 с.

[15]. А. Зюко, Д. Кловский, В. Коржик, *Теория электрической связи: учебник для вузов*, М.: Радио и связь, 1999, 432 с.

[16]. О. Голубничий, "Синтез систем корельованих сигналів з використанням доповненої процедури Грама-Шмідта", *Наукоємні технології*, Т. 40, № 4, С. 405-409, 2018.

[17]. D. Yu, L. Deng, "Gaussian Mixture Models", in *Automatic Speech Recognition. A Deep Learning Approach*, London: Springer-Verlag, 2015, Ch. 2, pp. 13-21. DOI: 10.1007/978-1-4471-5779-3_2.

[18]. A. Dempster, N. Laird, D. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1-38, 1977.

[19]. M. Gupta, Y. Chen, "Theory and Use of the EM Algorithm", *Foundations and Trends® in Signal Processing*, vol. 4, no. 3, pp. 223-296, 2011. DOI: 10.1561/2000000034.

[20]. N. Vlassis, A. Likas, "A Greedy EM Algorithm for Gaussian Mixture Learning", *Neural Processing Letters*, vol. 15, pp. 77-87, 2002.

[21]. T. Huang, H. Peng, K. Zhang, "Model Selection for Gaussian Mixture Models", *Statistica Sinica*, vol. 27, pp. 147-169, 2017.

[22]. О. Бакаева, "Определение минимального объёма выборки", *Вестник Мордовского университета. Серия "Физико-математические науки"*, № 4, С. 111-114, 2010.

[23]. О. Голубничий, "Аналіз конфіденційності передавання інформації у системах DSSS за умов обмеженості систем використовуваних сигнально-кодових конструкцій", *Захист інформації*, Т. 20, № 4, С. 221-230, 2018.

[24]. А. Голубничий, Г. Конахович, "Мультипликативно комплементарные бинарные сигнально-кодовые конструкции", *Известия высших учебных заведений. Радиоэлектроника*, Т. 61, № 10, С. 551-565, 2018.

[25]. О. Голубничий, "Синтез аналітичних форм опису автокореляційної функції узагальнених бінарних послідовностей Баркера типу 1 на основі її декомпозиції з використанням лінійних складових", *Наукоємні технології*, Т. 41, № 1, С. 10-15, 2019.

МЕТОД СТРУКТУРИЗАЦИИ КОДОВЫХ КОНСТРУКЦИЙ НА ОСНОВЕ ГАУССОВСКОЙ СМЕШАННОЙ МОДЕЛИ И ВЫДЕЛЕНИЯ ЕЁ КОМПОНЕНТ

Методы и подходы линейного криптоанализа криптографических алгоритмов направлены на анализ и обнаружение взаимосвязей между элементами открытого текста, шифротекста и ключа. В случае линейного криптоанализа псевдослучайных последовательностей и сигнально-кодовых конструкций, которые построены на их основе, анализу и обнаружению подлежат взаимосвязи между элементами этих последовательностей и сигнально-кодовых конструкций, а также между их внутренними структурами, их составляющими в образовываемых системах кодовых конструкций и т.д. Эффективность реализации линейного криптоанализа на разных его этапах может быть повышена при структуризации (обнаружении внутренних структур и взаимосвязей между ними) исследуемых кодовых конструкций, в отношении которых у стороны осуществления криптоанализа отсутствует любая априорная информация про их структуру, или кодовых конструкций, которые априори могут считаться имеющими стохастическую природу их образования. В статье предложен метод структуризации кодовых конструкций с априори неизвестной структурой на основе анализа корреляционных связей между кодовыми конструкциями, представляемыми в этом методе гауссовской смешанной моделью с последующим выделением её компонент и кластеризацией исследуемых кодовых конструкций с использованием обоснованной в статье параметрически-критериальной модификации ЕМ-алгоритма с удалением компонент. Метод позволяет выделять группы кодовых конструкций с взаимосвязанными структурами и далее выделять эти взаимосвязанные структуры в явном виде, в чём может заключаться решение ряда задач линейного криптоанализа, связанных с обнаружением структур и взаимосвязей между ними. Показан пример реализации предложенного метода для структуризации бинарных псевдослучайных последовательностей Баркера, которые используются в качестве сигнально-кодовых конструкций в широкополосных системах передачи информации и для которых из литературных источников известно, что они были синтезированы методом направленного перебора и поэтому имеют стохастический априори неструктурированный характер.

Ключевые слова: линейный криптоанализ, структуризация кодовых конструкций, обнаружение структур, обнаружение взаимосвязей, гауссовская смешанная модель, ЕМ-алгоритм, псевдослучайные последовательности, машинное обучение.

METHOD OF STRUCTURING CODE CONSTRUCTIONS BASED ON THE GAUSSIAN MIXTURE MODEL AND SEPARATION OF ITS COMPONENTS

Methods and approaches of the linear cryptanalysis of cryptographic algorithms are aimed at analyzing and detecting interconnections between plaintext, ciphertext and key elements. In the case of linear cryptanalysis of pseudorandom sequences and signal-code constructions based on them, the interconnections between elements of these sequences and signal-code constructions, as well as between their internal structures and their components in systems of code constructions, etc., are subject to analysis and detection. The effectiveness of linear cryptanalysis at its

different stages can be enhanced by structuring (detecting internal structures and interconnections between them) code constructions, in respect of which there is no a priori information about their structure, or code constructions, which a priori can be considered as constructions with a stochastic nature of their formation. The method of structuring code constructions with a priori unknown structures, which based on an analysis of cross-correlations between code constructions that are represented in this method by the Gaussian mixture model with a further separation of its components and clustering code constructions by means of modification (parametric and criteria features) of the EM-algorithm with removing components, is suggested in the article. The method allows selecting groups of code constructions with interconnected structures and then to detect these interconnected structures in an explicit form, which can be the solution of a number of problems of linear cryptanalysis related to the detection of structures and interconnections between them. An example of implementation of the proposed method for the structuring of binary pseudo-random Barker sequences, which are used as signal-code constructions in spread-spectrum telecommunications and were synthesized by the direct search method (as is known according to literary sources), and therefore have a stochastic a priori unstructured character, is shown in the article.

Keywords: linear cryptanalysis, structuring code constructions, detection of structures, detection of interconnections, Gaussian mixture model, EM-algorithm, pseudorandom sequences, machine learning.

Голубничий Олексій Георгійович, кандидат технічних наук, доцент, доцент кафедри телекомунікаційних систем Національного авіаційного університету.
E-mail: a.holubnychyi@nau.edu.ua.
Orcid ID: 0000-0001-5101-3862.
Голубничий Алексей Георгиевич, кандидат технических наук, доцент, доцент кафедры телекоммуникационных систем Национального авиационного университета.
Holubnychyi Alexei, PhD in Eng., Associate Professor at the Department of Telecommunication Systems, National Aviation University (Kyiv, Ukraine).

DOI: 10.18372/2410-7840.21.13766
УДК 004.056:004.75

ДЕКОМПОЗИЦІЙНА МОДЕЛЬ ПРЕДСТАВЛЕННЯ СМИСЛОВИХ КОНСТАНТ ТА ЗМІННИХ ДЛЯ РЕАЛІЗАЦІЇ ЕКСПЕРТИЗ У СФЕРІ ТЗІ

Олександр Корченко, Анатолій Давиденко, Максим Шабан

Проведення державних експертиз – це процес довготривалий і пов'язаний з можливими помилками як на етапі проведення проектних робіт, так і під час проведення самої експертизи. Тому актуальним науковим завданням є створення інформаційної системи, яка б допомагала експерту при побудові вихідних документів, а також дозволяла б експерту перевірити функціональний профіль захисту (ФПЗ). В роботі запропонована декомпозиційна модель, яка за рахунок сформованих множин вхідних та вихідних документів р-го проекту, а також множини смислових блоків, смислових констант та змінних р-го проекту дозволяє автоматизувати процес ідентифікації функціонального профілю захисту. Для цього ми провели декомпозицію вихідних документів з урахуванням етапів формування множин документів експертизи, смислових блоків та смислових змінних вихідних документів. Проаналізували принципи організації структури документів, які створюються на етапі державної експертизи комплексних систем захисту інформації (КСЗІ). Ввели поняття смисловий блок, смислова константа, смислова змінна. Смисловий блок – це стала семантична конструкція, яка має закінчене смислове значення. Смислова константа – це стійка смислова конструкція, час існування якої виходить за рамки проведення державної експертизи КСЗІ. В свою чергу, смислова змінна – це смислова конструкція, час існування якої дорівнює часу проведення державної експертизи КСЗІ. Все це дозволило прискорити процес створення вихідних документів державних експертиз КСЗІ. Розвитком даних робіт є розробка методу ідентифікації функціонального профілю захисту. Це дозволить формалізувати вимоги нормативного документу щодо його властивостей, що буде зроблено в подальших статтях.

Ключові слова: *державна експертиза КСЗІ, модель декомпозиції вихідних документів, грід-засоби, шаблон вихідних документів, систем підтримки прийняття рішень, смислові блоки, смислові константи, смислові змінні.*

Вступ. Проведення державних експертиз – це процес довготривалий і пов'язаний з можливими помилками як на етапі проведення проектних робіт, так і під час проведення самої експертизи. Експерт повинен опрацювати усі документи, які були розроблені на етапі проектних робіт і виходячи з отриманої інформації розробити групу вихідних документів, а саме: «**Програма та методика проведення експертизи**», «**Перелік тестів**», «**Протокол випробувань**», «**Експертний**