КОМП'ЮТЕРНИЙ ПРАКТИКУМ №2
З КУРСУ
МЕТОДИ КРИПТОАНАЛІЗУ 1

Статистичні критерії на відкритий текст

1 Мета роботи

Засвоєння статистичних методів розрізнення змістовного тексту від випадкової послідовності, порівняння їх, визначення похибок першого та другого роду.

2 Необхідні теоретичні відомості

Для побудови атак на криптосистеми необхідно аналізувати всі її частини. Вид шифротексту залежить від заданого відкритого тексту та криптографічного перетворення, яке відбулось над ним. Таким чином, статистичні властивості природних мов (якщо розглядати їх як відкриті тексти) будуть певним чином відображатися в шифротекстах. В даній лабораторній роботі розглядається задача розрізнення змістовних відкритих текстів від випадкової послідовності символів алфавіту, в ролі якої буде виступати результат перетворення відкритого тексту.

2.1 Математичне формулювання задачі розрізнення

Означення 1. *Алфавітом* Z_m будемо називати множину з m символів. Послідовність довжини L виду:

$$X = x_1 x_2 \dots x_L,\ x_i \in Z_m,\ i = \overline{1, L}$$

будемо називати *текстом* з алфавіту Z_m. Підпослідовність тексту довжини l будемо називати *l-грамою*.

Для природних мов характерно, що кожна l-грама має певну ймовірність появи в тексті. Також характерно, що після певної l-грами зустрічається інша l-грама з певною ймовірністю. В даній лабораторній роботі *відкритим текстом* будемо називати такий текст, який є змістовним.

Для постановки *задачі розрізнення* сформулюємо дві гіпотези відносно тексту X:

1. гіпотеза H_0 — текст X є відкритим текстом (змістовним текстом) з символів Z_m;
2. гіпотеза H_1 — текст X є випадковою послідовністю символів з Z_m.

Задача: побудувати статистичний критерій, який розрізняє гіпотези H_0 та H_1.

Для оцінки якості критерію будемо використовувати такі характеристики:

- $\alpha = P(H_1|H_0)$ — ймовірність помилки 1-го роду (англ. *false positive*), ймовірність назвати відкритий текст випадковою послідовністю;
- $\beta = P(H_0|H_1)$ — помилка 2-го роду (англ. *false negative*), ймовірність прийняти випадкову послідовність за відкритий текст.

Також для оцінювання критеріїв можна порівнювати мінімальну довжину тексту L для отримання заданих помилок 1-го і 2-го роду і складність роботи алгоритму. Далі будуть наведені найпопулярніші критерії на відкритий текст та їх варіації.

2.2 Критерій заборонених l-грам та його варіації

Дана група критеріїв перевіряє такі комбінації l-грам, які у змістовному тексті або взагалі неможливі, або трапляються дуже рідко. Будемо називати їх *забороненими*. Для застосування критерію необхідно попередньо розрахувати частоти символів/біграм/триграм/l-грам і визначити множину $A_{prh} \subset (Z_m)^l$, яка містить h_p заборонених l-грам.

2.2.1 *Критерій 1.0*

0. Розрахувати частоти l-грам для мови над алфавітом Z_m. Визначити множину A_{prh} на основі розрахованих частот.
1. Перевірити L знаків послідовності X.
2. Якщо серед L знаків зустрічається будь-яка l-грама x: $x \in A_{prh}$, то приймається гіпотеза H_1. Інакше приймається H_0.

Розглянемо деякі узагальнення та модифікації даного критерію. Для усіх модифікацій кроки 0-1 будуть спільними (будемо їх опускати при описі модифікацій), а відрізнятись такі модифікації будуть лише правилами розрізнення гіпотез.

2.2.2 *Критерій 1.1*

2. Якщо серед L знаків зустрічається будь-яка l-грама x: $x \in A_{prh}$, то додати її до множини A_{ap} (A_{ap} ініціалізується як порожня множина).
3. Якщо $|A_{ap} \cap A_{prh}| \geq k_p$ при $1 < k_p \leq h_p$, то приймається гіпотеза H_1. Інакше приймається H_0.

Тут k_p виступає значенням порогу, який попередньо обирається криптоаналітиком. Очевидно, що якщо покласти $k_p = 1$, то отримаємо критерій 1.0.

Наступні критерії напряму використовують статистичні дані.

2.2.3 *Критерій 1.2*

2. Серед L знаків порахувати частоти f_x усіх l-грам x, $x \in A_{prh}$.
3. Якщо існує така l-грама x, $x \in A_{prh}$, для якої $f_x > k_x$, то приймається гіпотеза H_1. Інакше приймається H_0.

Тут k_x — частота l-грами x для мови над алфавітом Z_m, яка отримана на кроці 0 в результаті попередніх обчислень.

2.2.4 *Критерій 1.3*

2. Серед L знаків порахувати частоти f_x усіх l-грам x, $x \in A_{prh}$. Обчислити значення:

$$F_p = \sum_{x \in A_{prh}} f(x).$$

3. Якщо $F_p > K_p$, то приймається гіпотеза H_1. Інакше приймається H_0. Можна інтуїтивно здогадатись з попереднього визначення, що K_p обчислюється таким чином:

$$K_p = \sum_{x \in A_{prh}} k_x.$$

Зауваження. В критеріях 1.0-1.3 можна використовувати різні множини A_{prh}.

2.3 Критерій частих l-грам та його варіації

Наступна група критеріїв працює з такими l-грамами мови, які навпаки дуже часто зустрічаються в змістовних текстах. Як і для першої групи критеріїв, потрібно попередньо розрахувати частоти. На основі цих розрахунків вибирають множину $A_{frq} \cap (Z_m)^l$, яка містить h_f l-грам, що зустрічаються у мові найчастіше.

2.3.1 *Критерій 2.0*

0. Розрахувати частоти l-грам для мови над алфавітом Z_m. Визначити A_{frq} на основі розрахованих частот.

1. Перевірити L знаків послідовності X.

2. Якщо існує така l-грама x: $x \in A_{frq}$, яка відсутня серед L знаків послідовності X, то приймається гіпотеза H_1. Інакше приймається H_0 (тобто, коли усі часті l-грами присутні в тексті).

Розглянемо модифікації даного критерію. Кроки 0-1 є однаковими для усіх модифікацій критерію частих l-грам.

2.3.2 *Критерій 2.1*

2. Якщо серед L знаків зустрічається будь-яка l-грама x: $x \in A_{frq}$, то додати її до множини A_{af} (A_{af} ініціалізується як порожня множина).

3. Якщо $|A_{af} \cap A_{frq}| \leq k_f$, $1 \leq k_f < h_f$, то приймається гіпотеза H_1. Інакше приймається H_0.

Аналогічно попередній групі критеріїв, k_f буде порогом, який попередньо обирається. Зазначимо, що якщо $k_f = h_f$ то отримаємо критерій 2.0.

Тепер розглянемо критерії з використанням статистичних даних.

2.3.3 *Критерій 2.2*

2. Серед L знаків порахувати частоти f_x усіх l-грам x, $x \in A_{frq}$.

3. Якщо існує така l-грама x, $x \in A_{frq}$, для якої $f_x < k_x$, то приймається гіпотеза H_1. Інакше приймається H_0 (тобто коли усі $f_x \geq k_x$).

2.3.4 *Критерій 2.3*

2. Серед L знаків порахувати частоти f_x усіх l-грам x, $x \in A_{frq}$. Обчислити таке значення:

$$F_f = \sum_{x \in A_{frq}} f_x.$$

3. Якщо $F_f < K_f$, то приймається гіпотеза H_1. Інакше приймається H_0. Тут:

$$K_f = \sum_{x \in A_{frq}} k_x.$$

Зауваження. Аналогічно критеріям заборонених l-грам, в критеріях 2.0-2.3 можна обирати різні A_{frq}.

2.4 Ентропійний критерій та критерій через розрахунок індексу відповідності

Зазвичай у криптографії будують модель джерела відкритих текстів*. Якщо розподіл імовірностей для l-грам $P(x_1x_2 \dots x_l)$ відомий, то можна розрахувати *питому ентропію на символ джерела* за формулою:

$$H_l = \frac{1}{l} \left(- \sum_{x \in (Z_m)^l} P(x) \cdot \log_2 P(x) \right).$$

Для природних мов ентропія на символ є константною характеристикою і може використовуватися для розпізнавання мови.

2.4.1 *Критерій 3.0*

0. Розрахувати частоти l-грам для мови над алфавітом Z_m. Розрахувати питому ентропію на символ джерела H_l для мови.

1. Серед L знаків послідовності X розрахувати частоти f_x усіх l-грам $x \in (Z_m)^l$. Розрахувати ентропію на символ джерела для послідовності $X - H'_l$.

2. Якщо $|H_l - H'_l| > k_H$, то приймається гіпотеза H_1. Інакше приймається H_0.

Такий критерій перевіряє, чи буде обчислена ентропія H'_l знаходитись в інтервалі $[H_l - k_H, H_l + k_H]$. При цьому поріг k_H обирається криптоаналітиком.

Ще однією характеристикою мов є *індекс відповідності*, який розраховується за такою формулою:

$$I_l = \frac{1}{L(L-1)} \sum_{x \in (Z_m)^l} c_x(c_x - 1),$$

де c_x – кількість l-грам, які зустрілись у послідовності.

2.4.2 *Критерій 4.0*

0. Розрахувати частоту l-грам для мови над алфавітом Z_m. Розрахувати індекс відповідності I_l.

1. Серед L знаків послідовності X розрахувати частоти c_x усіх l-грам $x \in (Z_m)^l$. Розрахувати індекс відповідності для послідовності $X - I'_l$.

2. Якщо $|I_l - I'_l| > k_I$, то приймається гіпотеза H_1. Інакше приймається H_0.

Аналогічно попередньому критерію, враховуючи поріг k_I, криптоаналітик будує інтервал $[I_l - k_I, I_l + k_I]$, в який має потрапити розрахований індекс відповідності I'_l.

*Розглянуто у курсі «Симетрична криптографія»

2.5 Критерій порожніх ящиків

Критерій порожніх ящиків оснований на відомій комбінаторній задачі про випадкове розміщення частинок по ящиках: нехай n частинок незалежно один від одного кидають в один з N ящиків, який розподіл має величина μ, що дорівнює кількості порожніх ящиків після розміщення всіх частинок? Модифікуючи класичний критерій порожніх ящиків, що перевіряє гіпотезу про вигляд розподілу деякої випадкової величини[†], створено два критерії на відкритий текст.

2.5.1 *Критерій 5.0*

0. Розрахувати частоти l-грам для мови над алфавітом Z_m.
1. Обрати серед l-грам, які зустрічаються найрідше в мові, j штук (для біграм $j = 50, 100, 200$). Позначимо множину обраних l-грам як B_{prh}. Кожному елементу з множини B_{prh} ставиться у відповідність один ящик.
2. Якщо серед L знаків послідовності X зустрічається l-грама з множини B_{prh}, то у відповідний їй ящик додається одна «частинка». Таким чином, кількість частинок у ящику є частотою зустрічей відповідної l-грами у послідовності X.
3. Обрахувати кількість порожніх ящиків, тобто тих l-грам з множини B_{prh}, які жодного разу не зустрілись у послідовності X. Позначимо кількість порожніх ящиків як f_{empt}.
4. Якщо $f_{empt} \leq k_{empt}$, то приймається гіпотеза H_1. Інакше приймається H_0.

Параметр k_{empt} є пороговим значенням, що обирається перед застосуванням критерію.

Наступний критерій є модифікацією критерію 5.0 шляхом аналізу кількості частих l-грам, що зустрічаються в тексті.

2.5.2 *Критерій 5.1*

0. Розрахувати частоти l-грам для мови над алфавітом Z_m.
1. Обрати серед найбільш частих l-грам j штук (для біграм $j = 50, 100, 200$). Позначимо множину обраних l-грам як B_{frq}. Кожному елементу з множини B_{frq} ставиться у відповідність один ящик.
2. Якщо серед L знаків послідовності X зустрічається l-грама з множини B_{frq}, то у відповідний їй ящик додається одна «частинка». Таким чином, кількість частинок у ящику є частотою зустрічей відповідної l-грами у послідовності X.
3. Обрахувати кількість порожніх ящиків, тобто тих l-грам з множини B_{frq}, які жодного разу не зустрілись у послідовності X. Позначимо кількість порожніх ящиків як f_{empt}.
4. Якщо $f_{empt} \geq k_{empt}$, то приймається гіпотеза H_1. Інакше приймається H_0.

Аналогічно до попереднього критерію, параметр k_{empt} є пороговим значенням, що обирається перед застосуванням критерію.

2.6 Структурний критерій

1. Згенерувати випадкову послідовність Z довжини L з алфавіту Z_m.
2. Обрати алгоритм стиснення даних (на власний вибір), наприклад, LZMA, що використовується в архіваторі `7-zip` та його модифікації, DEFLATE (архіватор `ZIP` та інші), BWT (архіватор `BZIP2`) та інші.

[†]Розглянуто у курсі «Методи теорії надійності та ризику»

3. Застосувати згенерований алгоритм стиснення до послідовностей Y (спотворений осмислений текст) та Z (випадкова послідовність символів алфавіту).

4. Оцінюючи результати стиснення, сформувати критерій на відкритий текст.

Додаткове завдання #1: Реалізувати один або кілька алгоритмів стиснення та, використовуючи власну реалізацію, побудувати критерій на відкритий текст. Кількість додаткових балів за кожен реалізований алгоритм стиснення: 2 бали.

Додаткове завдання #2: Обрати кілька алгоритмів стиснення та застосувати їх для побудови критерію. Порівняти результати критерію для різних алгоритмів стиснення. Кількість додаткових балів за кожен додатковий алгоритм стиснення: $0,5$.

3 Порядок виконання роботи і методичні вказівки

-1. Ознайомитись з порядком виконання комп'ютерного практикуму та відповідними вимогами до виконання роботи.

0. Уважно прочитати необхідні теоретичні відомості до комп'ютерного практикуму.

1. Створити новий репозиторій в системі контролю версій `Git` (бажано використовувати вебсервіс `GitHub` [‡]). Важливо:

   (а) репозиторій створюється перед початком роботи над програмним кодом (якщо репозиторій приватний, то перед початком роботи має бути надано доступ викладачу до даного репозиторію);

   (б) весь процес створення програмного коду має бути відображений у відповідних комітах проекту (для кожної атомарної функціональної зміни коду має бути власний коміт);

   (в) програмна реалізація не допускається до захисту при недотриманні вищевизначених вимог.

2. На великому тексті українською мовою (>1MB), де:

   (а) літера «ґ» замінена на літеру «г»;

   (б) видалений символ апострофу та усі інші спецсимволи в тексті, включно з пробілами;

   (в) текст містить лише маленькі літери алфавіту.

   необхідно розрахувати частоти літер і біграм, а також ентропію та індекс відповідності.

3. Отримати N текстів X українською мовою для довжин $L = 10$, 100, 1000 та 10000, для кожного з яких згенерувати спотворені тексти Y. Число N визначається відповідно до такої таблиці.

L	N
10 100 1000	10000
10000	1000

   Спотворення тексту виконується такими способами:

   (а) шляхом застосування шифру Віженера з *випадковим* ключем довжини $r = 1, 5, 10$:

$$y_i = \left(x_i + Key_{(i \bmod r)}\right) \bmod m;$$

[‡]Використання інших сервісів необхідно попередньо узгодити з викладачем

(б) шляхом застосування шифру афінної та афінної біграмної підстановки з *випадковими* ключами:

$$y_i = (a \cdot x_i + b) \bmod m^l,$$

де $a, b \in (Z_m)^l$ — ключі;

(в) y_i — рівномірно розподілена послідовність символів з $(Z_m)^l$;

(г) y_i обчислюється відповідно до такого співвідношення:

$$y_i = (s_{i-1} + s_{i-2}) \bmod m^l,$$

де $s_0, s_1 \in_R (Z_m)^l$.

4. Реалізувати критерії (відповідно до варіанту + структурний) і перевірити їх роботу на згенерованих N текстах для кожної довжини L. Розрахувати ймовірності похибок першого і другого роду.

Номер варіанту	*Критерії*
Парний	1.0-1.3, 3.0, 5.1
Непарний	2.0-2.3, 4.0, 5.0

Усі вищезгадані критерії (та інші формули), які використовували значення l, мають приймати значення $l = 1$ та $l = 2$, тобто реалізувати символьний та біграмний критерії.

5. Згенерувати випадковий текст довжини $L = 10000$, який точно не є зв'язним текстом українською мовою (наприклад, текст, який складається з величезної кількості літер а: *«аааааааа . . . »*). Застосувати один з варіантів спотворення (на вибір) до цього тексту, після чого застосувати один з реалізованих критеріїв (на вибір). Порівняти результати застосування критерію до різних текстів.

6. Оформити звіт до комп'ютерного практикуму.

Додаткове завдання: Реалізувати інші (не описані в комп'ютерному практикумі) критерії на відкритий текст. У такому випадку звіт має містити, окрім таблиці з ймовірностями похибок першого та другого роду, теоретичний опис даного критерію. За кожен додатковий реалізований критерій до +1 балу до рейтингу.

Комп'ютерний практикум виконується у такому ж складі бригади, як виконувався комп'ютерний практикум №1. Зміна складу бригади та способу виконання роботи протягом семестру можлива лише при узгодженні цього з викладачем комп'ютерних практикумів.

4 Оформлення звіту

Звіт про виконання комп'ютерного практикуму оформлюється згідно зі стандартними правилами оформлення наукових робіт за допомогою системи набору і верстки LaTeX, причому дозволяється використовувати розмір шрифту 12pt та одинарний міжрядковий інтервал. Звіт обов'язково має містити:

- мету комп'ютерного практикуму;
- постановку задачі та варіант завдання;
- хід роботи;
- опис множин заборонених/частих символів, які було отримано при виконанні завдання;
- окремі таблиці для кожного зі способів спотворення, які містять ймовірності похибок першого та другого роду для кожного з критеріїв, що реалізовуються в роботі, для різних значень L та l; шаблон таблиці:

Спотворення за допомогою шифру Віженера ($r = 5$)					
L	*Номер критерію*	**FP** ($l = 1$)	**FN** ($l = 1$)	**FP** ($l = 2$)	**FN** ($l = 2$)
10	2.0 (обрані порогові значення)				
	2.1				
	2.2				
	2.3				
	4.0				

- для кожного критерію вказати значення порогових значень, що було підібрано при реалізації (наприклад, в дужках в таблиці біля номера критерію, як вказано в шаблоні таблиці);
- опис алгоритму стиснення, що був обраний для розробки структурного критерію;
- опис запропонованого структурного критерію, що базується на основі результатів стиснення;
- опис труднощів, що виникали при виконанні комп'ютерного практикуму, та шляхи їх розв'язання;
- висновки (аналіз ефективності реалізованих критеріїв, порівняння їх між собою, порівняння результатів для різних значень L, r тощо).

Лістинги програми дозволяється не включати у звіт.

5 Порядок захисту комп'ютерного практикуму

Для зарахування комп'ютерного практикуму студенту необхідно виконати захист теоретичної та практичної частин роботи (за умови своєчасного надання доступу викладачеві до `Git`-репозиторію, що містить код програми). Студент має можливість здавати теоретичну та практичну частини комп'ютерного практикуму в різні дні в довільному порядку.

6 Контрольні питання

1. Задача розрізнення гіпотез, помилки першого та другого роду.
2. Критерій заборонених l-грам та його варіації.
3. Критерій частих l-грам та його варіації.
4. Ентропійний критерій та критерій через розрахунок індексу відповідності.
5. Критерій порожніх ящиків.
6. Структурний критерій.

Оцінювання комп'ютерного практикуму

Можлива кількість рейтингових балів 12

- Програмна реалізація 5
- Теоретичний захист роботи 6
- Своєчасне виконання роботи 1

Несвоєчасне виконання роботи ..... -1 бал за кожен тиждень пропуску

Академічний плагіат ..... -10 балів до рейтингу з вимогою виконати комп'ютерний практикум повторно та без можливості складання іспиту на основній сесії