УДК 004.93
В.В. Томчук
в/ч А3346
г. Севастополь, ул. Соловьева 1, 99029

СПОСОБ РЕШЕНИЯ ЗАДАЧИ СНИЖЕНИЯ РАЗМЕРНОСТИ В АНАЛИЗЕ ШИФРОВАННЫХ СООБЩЕНИЙ

Излагаются результаты разработки методики решения «Первой задачи» криптографического анализа шифрованных сообщений.

За последние два десятилетия информационная безопасность приобрела роль одного из определяющих факторов в обороноспособности всех государств мира. Статистика итогов каждого года свидетельствует о росте мировой компьютерной преступности и убытков, ставших следствием утрат информации. В сложившихся условиях важным является интенсификация внешней и внутренней стратегий государственной информационной безопасности.

Благодаря появлению и развитию глобальной компьютерной сети объемы информационных потоков резко возросли, требования к оперативности и качеству их аналитической обработки возрастают соответственно. Сформулируем перечень требований к средствам аналитической обработки информационных объектов, которые составляют поток:

1) универсальность;
2) минимизация времени выполнения процесса анализа и принятия решения;
3) минимизация объема требуемого машинного ресурса;
4) автоматизация общей процедуры анализа.

В условиях высоких скоростей и больших объемов передачи информации существующие решения задачи анализа и классификации шифрованных сообщений не в полной мере удовлетворяет перечисленным требованиям, и как следствие, пока не являются удовлетворительными. Более того, возникают определенные трудности при решении той же задачи, в условиях распределенной передачи шифрованных сообщений. Эта разновидность передачи сообщения характерна тем, что криптограмма делится на некоторое количество частей и передача получателю производится различными маршрутами (трафиками). Иными словами, максимум того, что может быть доступно для проведения анализа – часть шифрованного сообщения. Следует отметить, что в этом случае происходит значительное снижение общей сопутствующей информации [1], которую можно извлечь. Для получения объективного представления о специфике исследуемой криптограммы, как части информационного потока, назовем причины препятствующие проведению операции предварительной оценки сообщения и снижающие эффективность криптоаналитической структуры (эффект отвлечения внимания):

1) протокол оформления криптограммы изначально не предусматривает необходимого (для имеющихся способов оценки) количества информационных признаков;
2) аппаратное шифрование производится без явных «неизбежных меток»;
3) криптограмма содержит заведомо ложные признаки;
4) шифртекст криптограммы является хаотическим, неупорядоченным набором символов (муляж, лже-сообщение);
5) для пересылки криптограммы применен метод распределенной передачи.

Принимая во внимание перечисленное, необходимо найти такой способ обработки шифрованных сообщений (шифрованных текстов), который позволит компенсировать сложность, определяемую пунктами 1-5. В качестве такого способа предлагается методика обработки и классификации шифрованных сообщений (шифрованных текстов), которая формально базируется на концепции снижения размерности и многомерного статистического анализа [2].

Исходными данными на этапе исследования выступает поток отдельных сообщений

$$T = (T_1, ..., T_Q),$$

где T_q – выбранное сообщение, $q \in [1, \ldots, Q]$.

Поток сообщений рассматривается в двух состояниях, первое — сообщения T_q открытые, второе — сообщения T_q зашифрованные. Зашифрованные сообщения предполагают специфическую форму своего представления в виде шифрованного текста (ШТ).

Пусть дано множество M сообщений T_q. Множество M объединено завершенностью логического оформления этих сообщений в дальнейшем называемых допустимыми или регулярными. Множество M допустимых сообщений покрыто конечным числом L подмножеств сообщений, выработанных с использованием различных способов зашифрования $K_1, \ldots, K_L$,

$$M = \bigcup_{l=1}^{L} K_l ,$$

где подмножество K_l называют классом l-признака преобразования (или просто классом).

Разбиение M не определено. Задана лишь некоторая информация I_0 о классах K_1, …, K_L в виде n-выборок умышленно выработанных сообщений T', принадлежность каждого из которых к своему классу не вызывает сомнений, т.е.

$$M = [K_1 \in [T'_1,...,T'_m], \quad ..., \quad K_L \in [T'_1,...,T'_r]],$$

где $T_i \in M$ — допустимый текст, который определен значениями некоторых характеристик q_i.

Совокупность заданных значений q_i определяет описание сообщения или его образ $I(T)$

$$I(T) = \{q_i\}, \quad q_i \in [0, \quad ..., \quad \infty].$$

Основная задача образного представления символьной информации состоит в том, чтобы создать такое описание допустимого сообщения $I(T)$ и информации о классах $I_0(K_1, …, K_L)$, чтобы в последствии ее кластеризовать, т.е. вычислить значение предиката $P_l(T)$, P_l : $[T \in K_l]$. Информацию I_0 принято называть обучающей, предикаты $P_l(T)$ – элементарными предикатами.

Решение этой задачи следует искать в структуре исследуемого сообщения.

Множество ШТ, выработанных соответствующим криптографическим алгоритмом, хорошо описывается моделью метаязыка. В терминах математической теории контекстно-свободных языков [3]:

множество V_T – терминальных символов, словарь или лексикон;
множество V_N – синтаксических переменных;
множество $\hat{A}$ правил преобразования, каждое из которых имеет вид V_N *→ V * (V=$V_T \cup V_N$) } (1)

Заметим что, рассматриваемый метаязык, способный порождать собственные высказывания, заранее наделяется набором известных характеристик (1). Но работа любого криптоалгоритма, в значительной мере, упрощена по сравнению с языками машинного программирования, и тем более – естественными языками.

Замечание 1. Адаптируя метаграмматическую систему, специфичность элементов множества V_N будет делать общий вид модели ШТ, в некотором смысле примитивным. Это замечание будет подкреплять точку зрения на систему правил образования продуктов метаязыка.

Прокомментируем приведенное замечание. Специфичность множества V_N обусловлена вследствие отсутствием глобального [3] синтаксического свойства. Теоретически возможна [1] межсимвольная сериальная, последовательная корреляция в выходном потоке, что составляет синтаксическую независимость [3].

Обозначим порядок применения правил (1). Для двух потоков a и b, принадлежащих V*, можно записать $a \rightarrow b$, если существуют элементы α, β, x, y, такие что $a = x\ \alpha\ y$ и $b = x\ \beta\ y$ и отношение $\alpha \rightarrow \beta$ принадлежит множеству $\hat{A}$.

Очевидно, что ШТ составляется из некоторого множества символов, которое трактуется алфавитом зашифрования, а криптоалгоритм производит перестановки по определенным правилам. Таким образом

$$T = (A_{\Sigma^*}, \Lambda_{\Sigma^*}), \qquad (2)$$

где $A_{\Sigma^*} = \mathbf{U}_{n=1}^{N}\left(\sum_{k=0}^{M} x_m p^k\right)_n$ – алфавит зашифрования; $\Lambda_{\Sigma^*} = (\gamma_C, \gamma_Б, \gamma_{Сл}, \gamma_{Сег}, \gamma_И)$ – комплексное выражение набора правил преобразования метаязыка Σ*; γ_C – символьная группа правил; $\gamma_Б$ – блочная группа правил; $\gamma_{Сл}$ – словарная группа правил; $\gamma_{Сег}$ – сегментарная группа правил;
$\gamma_И$ – иерархическая группа правил.

Определим грамматический контур метаязыка Σ*. Каждая из групп правил Λ_{Σ^*} может быть определена следующим образом.

1. Символьная группа правил γ_C. Каждую элементарную текстовую единицу (символ) a_n [4] представим вектором, величина которого $|a_n|$, обусловлена ее весовой долей во всем объеме текста. Вектор $\vec{a}_n$ проецируется на оси полярной системы координат:

$$a_n = (|a_n|, \vec{a}_n); \qquad \vec{a}_n : (\varphi, |a_n|),$$

где φ – угол вектора $\vec{a}_n$.

Таким образом,

$$\gamma_C = \{\vec{a}_n\}. \qquad (3)$$

2. Блочная группа правил $\gamma_Б$ отражает принцип формирования некоторой группы символов E'*. В блочных алгоритмах шифрования этими группами символов являются S-блоки [1]:

$$\gamma(E'^*) : \vec{E'}^* = \sum_{n=1}^{N} \vec{a}_n; \qquad \gamma_Б = \gamma(E'^*); \qquad (4)$$

$$E'^* = E'^*_2 \subset E'^*_1,$$

где E'_2* и E'_1* – изменяемая и основная (неизменяемая) части S-блока.

3. Словарная группа правил $\gamma_{Сл}$ определяется всем разнообразием возможных вариантов формирования S-блоков криптоалгоритма, т.е.

$$\gamma_{Cл} = G\left[r, l, \sigma, \sigma^{-1}\right], \quad (5)$$

где G – набор всех вариантов формирования S-блоков; r, l – правая и левая части S-блока; σ – сложение с переносом (правый циклический сдвиг) [1]; σ^{-1} – вычитание с переносом (левый циклический сдвиг) [1]

$$\left.\begin{array}{l}\sigma \ : x \to y \| y \| = (\| x \| + 1) \\ \sigma^{-1} : x \to y \| y \| = (\| x \| - 1)\end{array}\right\} \text{ (мод. } 2^N\text{)},$$

где N – размер S-блока.

4. Сегментарная группа правил $\gamma_{Ceг}$ для модели метаязыка $\gamma_{Ceг}$ определена быть не может, так как множеству синтаксических переменных V_N (1) свойственна локальность синтаксических свойств (замечание 1 и комментарий).

5. Иерархическая группа правил $\gamma_{И}$ организует структуру исследуемой модели шифрованного сообщения, которая состоит из текстовых единиц t различных уровней b [4]:

I уровень – символ;

II уровень – блок;

III уровень – текст.

Наличие синтаксической независимости, которая свойственна контекстно-свободным языкам [3], обуславливает отсутствие грамматического механизма вывода многоуровневых комбинаций символов. Иными словами, зависимое сочетание блоков, которые образуют группы, исключается в принципе. Для взвешивания текстовой единицы t^b во всем объеме ШТ T определяется ее соответствием одноименному коэффициенту k^b [4], значения которого приведены в таблице 1 (см. столбец 3).

Таблица 1 – Таблица соответствия весовых коэффициентов текстовым единицам

№ t^b	Название текстовой единицы	Коэффициент k^b
1	2	3
1	символ	0.146
2	блок	0.236
3	текст	1.000

Суммарный коэффициент текстовой единицы k^{Σ} является результатом суперпозиции

$$a^b = \sum_{j=1}^{b} \sum_{i=1}^{I} k_i^j \left|\mathbf{a}\right|_i .$$

Формализация иерархической группы правил имеет следующее выражение:

$$\gamma_{И} = F(t^b) . \quad (6)$$

Таким образом, описанные группы правил γ_i и множество терминальных символов A_{Σ^*} интерпретируют состав формулы 2, определяя модель ШТ.

Рассмотрим синтез образа шифрованного текста. Построение описания информационного объекта $1(T)$ приобретает форму изображения конфигурации.

Определение 1. Изображением ШТ $1(T)$ – называется одномерный массив (g_1, ... g_H), $I(T) = (g_1, ..., g_H) = \{g_h\} h \in H$, получаемый в результате преобразований

$$g_h = \sum_{j=1}^{J(H)} t(h)_j^2 , \quad (7)$$

где $t(h)_j^2$ – множество блоков J индекса h.

Индекс h вычисляется путем обработки блока в полярной системе координат с разрешением H, H=360/h [4]

$$h = \left[E'/H\right] . \quad (8)$$

Изображение ШТ $1(T)$ состоит из упорядоченного списка образующих-признаков g_h, а доминанта изображения d (максимальная образующая изображения) имеет высоту всего оперативно-визуального пространства L. Все остальные образующие имеют текущие высоты l_n и определяются в соотношением [4]:

$$l_h = g_h \times L/d . \quad (9)$$

Обработка изображения конфигурации объекта (ИКО) предполагает приведение его к необходимому качественному виду. Для этого могут быть использованы стандартные операции аппроксимации и сглаживания. Проведение операций обработки — суть измерение ряда параметров полученной огибающей:

1) коэффициент мощности термина $A_{H_n^j}$ и изображения A;
2) коэффициент структуры термина $B_{H_n^j}$;
3) коэффициент относительной частоты появления доминанты термина H_n^i;
4) коэффициент эффективной ширины изображения C;
5) среднее число пересечений D_j уровня x_m;
6) средняя длительность всплесков изображения E_j.

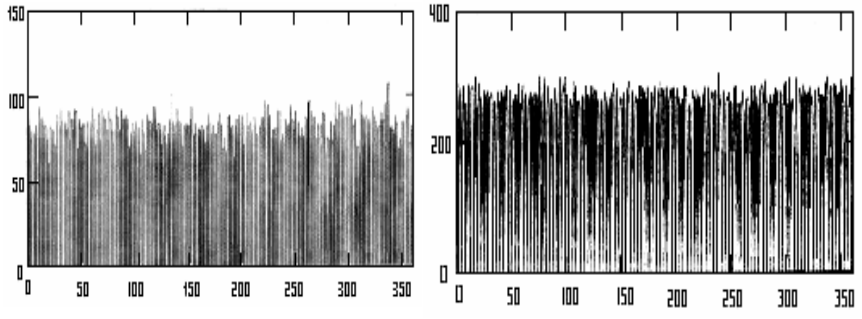
Определение 2. Термином H_n^j p-го порядка определим совокупность подконфигураций $c^i_{n(r-m)}$, объединенных в группу, которая образует фрагмент изображения. Фрагмент отождествляется признаком изображения, который в геометрическом смысле интерпретируется одной из простых функций. Порядок j термина определяется количеством входящих в него групп.

В пределах одного термина выделим максимальную образующую $d_{H_n^j}$ – доминанту H_n^j-термина, в пределах всего изображения – доминанту конфигурации d.

Предложенный перечень параметров составляет код информационного объекта:

$$X = \left\{A_I, A_{H_n^j}, B_{H_n^j}, S_{H_n^j}, C, D, E\{\zeta\}\right\}. \qquad (10)$$

На рисунке 1 представляет изображения двух информационных объектов, к которым были применены разные алгоритмы шифрования – «Idea» и «rc-2»

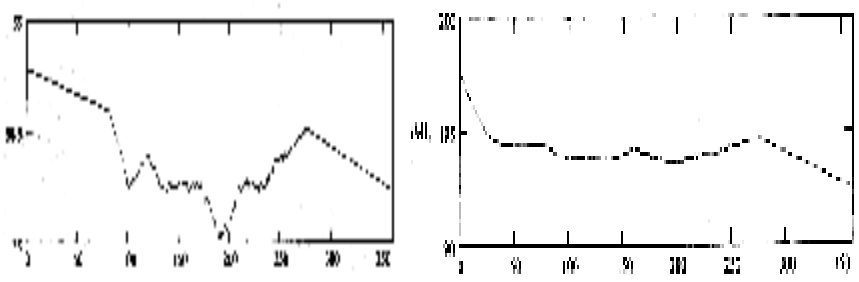


а — «Idea» б — «rc-2»

Рисунок 1 — Изображения информационных объектов, выработанных различными КА: а — «Idea»; б — «rc-2»

После осуществления качественного преобразования полученных изображений информационных объектов, огибающие принимают вид показанный на рисунке 2.

Предложенный способ (1) построения модели ШТ (2), интерпретируемой составом собственных



а — «Idea» б — «rc-2»

Рисунок 2 — Огибающие изображений информационных объектов, выработанных различными КА: а — «Idea»; б — «rc-2»

компонентов (3-6), имеет важную особенность – представление в виде изображения $1(T)$, которое производится в результате преобразований (7-9). Однако, ИКО является слишком избыточным описанием информационного объекта, в силу чего, возникает необходимость вывода наиболее информативного набора признаков, обеспечивающих снижение размерности исходного описания информационного объекта $I(T) = (g_1,..., g_H) = \{g_h\}$. Таким образом, выражение (10) является частным решением задачи снижения размерности при обработке шифрованных сообщений (шифрованных текстов).

Предложенный способ выделения кода информационного объекта (10), позволяет повысить качество классификации шифрованного сообщения (шифрованного текста) в обстановке большого разнообразия порождающих источников и минимальной степени визуализации признаков принадлежности информационного объекта к выработавшему его источнику.

Дальнейшее исследование предмета статьи предполагает разработку соответствующих методов распознавания информационных объектов, а также разработку системы классификации — процессора образа.

Библиографический список

1. Конхейм А. Г. Основы криптографии / А.Г. Конхейм — М.: Радио и связь, 1987. — 412 с.
2. Прикладная статистика. Классификация и снижение размерности / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин. — М.: Финансы и статистика, 1989. — 606 с.
3. Гренандер У. Лекции по теории образов. Синтез образов. Пер. с англ. / У. Гренандер. — М.: Мир, 1979. —382 с.

4. Манухін О. В. Можливості смислової обробки інформації за допомогою теорії розпізнавання зразків / О.В. Манухін // Захист інформації. — К., 1997. — № 1. — С. 169 – 171.

Поступила в редакцию 16.05.2005 г.