УДК 004.415.24; 519.22

Л.Л. НИКИТЕНКО

ИССЛЕДОВАНИЕ КРИТЕРИЯ СТОЙКОСТИ ПРИ ПАССИВНЫХ АТАКАХ

Приведено определение распределения, δ-приближенного относительно известного распределения. Для δ-приближенного распределения получены оценки стойкости стегосистемы для критерия стойкости при пассивных атаках. Также установлена связь между параметрами критерия Неймана–Пирсона и критерием стойкости для δ-приближенного распределения. Получено неравенство, позволяющее оптимизировать стойкость стегосистемы и уровень значимости критерия Неймана–Пирсона.

Введение. В работе рассматривается теоретико-информационная модель стегосистемы с пассивным нарушителем. Стойкость стегосистемы при пассивных атаках определяется невозможностью принятия решения о наличии или отсутствии встроенной информации сторонним наблюдателем на основе информации о распределении контейнеров и стего. Предполагается, что отправитель, получатель и нарушитель (в данном случае наблюдатель) владеют одинаковой информацией о распределении контейнеров и стего.

Рассмотрим стегосистему с множеством контейнеров C и множеством стего S. Будем считать, что $P_C(q)$ и $P_S(q)$ вероятности того, что изображение $q \in Q$, где Q – множество принятых и отправленных изображений, является пустым контейнером и стего соответственно. Критерий стойкости стего при пассивных атаках [1] основывается на понятии относительной энтропии. Стегосистема называется ε-стойкой, если относительная энтропия между вероятностным распределением контейнеров и стего не превышает ε:

$$D(P_C \,//\, P_S) = -\sum_{q \in Q} P_C(q) \log \frac{P_C(q)}{P_S(q)} \le \varepsilon \ . \quad (1)$$

В работе [2] этот критерий был распространен на конкретное изображение. Изображение рассматривалось как стегосистема, контейнерами которой являются непересекающиеся части изображения, например, отдельные пикселы или коэффициенты преобразованной матрицы цифрового изображения. В данной работе критерий (1) будет рассматриваться для обоих случаев.

1. δ-приближенные распределения

Определение 1. Распределение $P_X(q)$ будем называть δ-приближенным относительно распределения $P_0(q)$, если

$$\max_{q\in Q}\left|\frac{P_X(q)-P_0(q)}{P_0(q)}\right|\le \delta<1. \tag{2}$$

Теорема 1. Пусть распределение $P_X(q)$ случайной величины $q\in Q$ является δ-приближенным относительно распределения $P_0(q)$, тогда

$$D(P_0 \,//\, P_X)<\frac{\delta}{\ln a}. \tag{3}$$

Доказательство. Обозначим δ_q относительную разность вероятностей $P_X(q)$ и $P_0(q)$:

$$\delta_q=\frac{P_X(q)-P_0(q)}{P_0(q)}. \tag{4}$$

Тогда

$$P_X(q)=P_0(q)(1+\delta_q). \tag{5}$$

Относительная энтропия между вероятностным распределением $P_0(q)$ и $P_X(q)$

$$D(P_0 \,//\, P_X)=-\sum_{q\in Q}P_0(q)\log\frac{P_0(q)}{P_X(q)}=$$

$$=-\sum_{q\in Q}P_0(q)\log\frac{P_0(q)}{P_0(q)(1+\delta_q)}=\sum_{q\in Q}P_0(q)\log(1+\delta_q). \tag{6}$$

Для функции $\ln(1+x)$ справедливо неравенство [3]

$$\frac{x}{1+x}<\ln(1+x)<x,\qquad (x>-1). \tag{7}$$

Перейдем в (6) к натуральному логарифму и воспользуемся неравенством (7):

$$D(P_0 \,//\, P_X)<\frac{1}{\ln a}\sum_{q\in Q}P_0(q)\delta_q\le\frac{\delta}{\ln a}.$$

Таким образом,

$$D(P_0 \,//\, P_X)<\frac{\delta}{\ln a}.$$

Следствие 1. Чтобы стегосистема была ε-стойкой, достаточно, чтобы распределение стего было $\varepsilon\ln a$-приближенным относительно распределения контейнеров.

Из теоремы 1 и следствия 1 можно сделать вывод, что при встраивании сообщения в конкретное изображение в качестве оценки стойкости стегосистемы можно использовать неравенство (3). При встраивании сообщения с заданным по критерию уровнем стойкости ε достаточно выполнения условия

$$\max_{q\in Q}\left|\frac{P_S(q)-P_C(q)}{P_C(q)}\right| \le \varepsilon \ln a\,. \tag{8}$$

2. Связь между δ-приближением контейнеров, δ_s-приближением стего и критерием стойкости

Пусть из всего множества контейнеров C в качестве реальных выбрано множество контейнеров C_δ, δ-приближенных относительно распределения $P_0(q)$, которое известно и отправителю, и получателю. Встраивание секретного сообщения *m* из множества возможных сообщений *M* отправителем осуществляется так, что стего являются δ_s-приближенными относительно распределения $P_0(q)$. Исследуем взаимосвязь между δ, δ_s и $D(P_C // P_S)$. При этом используем равенство (5) и неравенство (7):

$$D(P_C // P_S) = -\sum_{q\in Q} P_C(q)\log\frac{P_C(q)}{P_S(q)} = -\sum_{q\in Q} P_C(q)\log\frac{P_C(q)P_0(q)}{P_S(q)P_0(q)} =$$

$$= -\sum_{q\in Q} P_C(q)(\log\frac{P_C(q)}{P_0(q)} + \log\frac{P_0(q)}{P_S(q)}) = \frac{1}{\ln a}\sum_{q\in Q} P_C(q)(\ln\frac{P_0(q)}{P_C(q)} - \ln\frac{P_0(q)}{P_S(q)}) =$$

$$= \frac{1}{\ln a}\sum_{q\in Q} P_0(q)(1+\delta_q)(\ln\frac{P_0(q)}{P_C(q)} - \ln\frac{P_0(q)}{P_S(q)}) =$$

$$= \frac{1}{\ln a}\sum_{q\in Q} P_0(q)(\ln\frac{P_0(q)}{P_C(q)} - \ln\frac{P_0(q)}{P_S(q)}) + \frac{1}{\ln a}\sum_{q\in Q} P_0(q)\delta_q(\ln\frac{P_0(q)}{P_C(q)} - \ln\frac{P_0(q)}{P_S(q)}) =$$

$$= D(P_0 // P_S) - D(P_0 // P_C) + D_\delta(P_0 // P_S) - D_\delta(P_0 // P_C). \tag{9}$$

В формуле (9) через $D_\delta(P_0 // P_S)$ и $D_\delta(P_0 // P_C)$ обозначены составляющие, обусловленные отклонениями вероятностей распределения контейнеров и стего от соответствующих вероятностей распределения $P_0(q)$.

Первые два слагаемых в (9) в соответствии с теоремой 1 и тем, что относительная энтропия всегда неотрицательна, ограничены. Следует оценить последние два слагаемые. Рассмотрим их подробнее.

$$D_\delta(P_0 // P_S) = -\frac{1}{\ln a}\sum_{q\in Q} P_0(q)\delta_q \ln\frac{P_0(q)}{P_S(q)} = \sum_{q\in Q} P_0(q)\delta_q \ln(1+\delta_{Sq}) <$$

$$< \frac{1}{\ln a}\sum_{q\in Q} P_0(q)\delta_q\delta_{Sq} \le \frac{\delta\delta_S}{\ln a}, \tag{10}$$

где

$$\delta_{Sq} = \frac{P_S(q) - P_0(q)}{P_0(q)}.$$

Аналогично получаем

$$D_\delta(P_0 \,//\, P_C) = -\frac{1}{\ln a}\sum_{q\in Q} P_0(q)\delta_q \ln\frac{P_0(q)}{P_C(q)} = \sum_{q\in Q} P_0(q)\delta_q \ln(1+\delta_q) <$$

$$< \frac{1}{\ln a}\sum_{q\in Q} P_0(q)\delta_q^2 \le \frac{\delta^2}{\ln a}. \qquad (11)$$

С другой стороны, используя (7) получаем нижнюю границу для $D_\delta(P_0 \,//\, P_C)$:

$$D_\delta(P_0 \,//\, P_C) = -\frac{1}{\ln a}\sum_{q\in Q} P_0(q)\delta_q \ln\frac{P_0(q)}{P_C(q)} = \frac{1}{\ln a}\sum_{q\in Q} P_0(q)\delta_q \ln(1+\delta_q) >$$

$$> \frac{1}{\ln a}\sum_{q\in Q} P_0(q)\frac{\delta_q^2}{(1+\delta_q)} \ge 0. \qquad (12)$$

Из формул (9), (10), с учетом неотрицательности относительной энтропии и неравенств (11) и (12), получаем верхнюю оценку $D(P_C \,//\, P_S)$, выраженную через значения δ и δ_s:

$$D(P_C \,//\, P_S) < \frac{1}{\ln a}\delta_S(1+\delta). \qquad (13)$$

Полученный результат сформулируем в виде теоремы.

Теорема 2. Если множество контейнеров состоит из контейнеров с распределением, δ-приближенным относительно распределения $P_0(q)$, а множество стего состоит из стего с распределением δ_s-приближенным относительно распределения $P_0(q)$, тогда для критерия стойкости имеет место соотношение

$$D(P_C \,//\, P_S) < \frac{1}{\ln a}\delta_S(1+\delta).$$

Следствие 2. Пусть при встраивании сообщения в изображение меняется параметр $z \in Z$, характеризующий непересекающиеся элементы изображения (пикселы, коэффициенты вейвлет-преобразования и т.д.). Пусть контейнером является изображение, в котором распределение параметра $z \in Z$ является δ-приближенным относительно известного распределения $P_0(z)$. Тогда по распределению параметра $z \in Z$ в полученном стего можно определить его δ_s-приближение относительно $P_0(z)$ и оценить стойкость стегосистемы, контейнерами в которой являются непересекающиеся элементы изображения, характеризующиеся параметром $z \in Z$.

3. Критерий стойкости при выполнении критерия Неймана–Пирсона

Кратко опишем критерий Неймана–Пирсона [4]. Рассмотрим задачу обнаружения наличия встроенной информации (например, цифрового водяного знака ЦВЗ) как статистическую задачу проверки основной гипотезы H_0 – «изображение не содержит ЦВЗ», при условии, что для нее имеется всего одна альтернатива, гипотеза H_1 – «изображение содержит ЦВЗ».

Всякое правило обнаружения характеризуется вероятностью принять ту или другую гипотезу в зависимости от наблюдаемого значения случайной величины. Пусть $\pi(Z)$ есть «критическая» вероятность – вероятность отвергнуть основную гипотезу, если наблюдаемое изображение есть Z, т. е.

$$\pi(Z) = P(H_1 \mid Z) \tag{14}$$

(вероятность принять гипотезу H_1 о наличии ЦВЗ в изображении при условии, что наблюдалось Z). Качество правила определяется вероятностями принятия и отвержения каждой из гипотез в зависимости от того, какая из гипотез верна. Его характеризуют вероятностями ошибок.

Ошибка первого рода – отвергнуть истинную основную гипотезу H_0, она обычно обозначается α:

$$\alpha = P(H_1 \mid H_0)\,. \tag{15}$$

Ошибка второго рода – принять основную гипотезу, если верна альтернатива, ее вероятность обозначается β:

$$\beta = P(H_0 \mid H_1)\,. \tag{16}$$

α – уровень значимости критерия, $1-\beta$ – его мощность. Уровень значимости соответствует вероятности ложной тревоги обнаружения ЦВЗ, мощность – это вероятность обнаружения ЦВЗ. Как правило, уровень значимости выбирается заранее, мощность критерия максимизируют, т.е. вероятность ошибки второго рода стараются сделать минимальной. Если выбрана критическая вероятность, то

$$\alpha = \int \pi(Z) F_0(dZ)\,, \quad \beta = \int (1-\pi(Z)) F_1(dZ)\,, \tag{17}$$

где $F_i(Z)$ – распределение случайной величины Z при гипотезе $H_i\,(i = 0,1)$.

Наиболее мощный критерий с заданным уровнем значимости – критерий Неймана–Пирсона. Он определен для случая, когда мера F_1 абсолютно непрерывна относительно меры F_0: для всех борелевских множеств $C \in R^m$

$$F_1(C) = \int_C f(Z) F_0(dZ)\,. \tag{18}$$

Рассмотрим два важных случая, когда равенство (18) выполнено.

1. Пусть F_1 и F_2 имеют плотности вероятностей относительно лебеговой меры в R^m, равные $f_1(Z)$ и $f_2(Z)$ соответственно, причем $f_0 > 0$. Тогда равенство (18) выполняется, если

$$f(Z) = \frac{f_1(Z)}{f_0(Z)}. \tag{19}$$

2. Пусть обе меры $F_i(dZ)$ дискретны, т.е. можно указать такую последовательность $Z_i \in R^m$, что $\sum F_0(\{Z_i\}) = \sum F_1(\{Z_i\}) = 1$ и $F_0(\{Z_i\}) > 0$ для всех Z_i. Тогда

$$f(Z_i) = \frac{F_1(\{Z_i\})}{F_0(\{Z_i\})} \tag{20}$$

(значения $f(Z)$ при $Z \neq Z_i$ роли не играют).

Пусть J_t – такое множество, что $f(Z) \geq t$ при $Z \in J_t$, $f(Z) \leq t$ при $Z \in R^m - J_t$. При $F_0(J_t) = \alpha$, критерий является наиболее мощным при уровне значимости α для непрерывного случая. Кроме того, всегда можно разделить R^m на два множества $D_{t\alpha}$ и $R^m - D_{t\alpha}$, такие, что $f(Z) \geq t_\alpha$ на $D_{t\alpha}$ и $f(Z) \leq t_\alpha$ на $R^m - D_{t\alpha}$, тогда $F_0(D_{t\alpha}) = \alpha$.

Для дискретного случая R^m делится на три множества $D_{t\alpha}$, где $f(Z) > t_\alpha$, Γ_α, где $f(Z) = \alpha$, и $C_{t_\alpha} = R^m - D_{t\alpha} \cup \Gamma_\alpha$. Если $\pi(Z)$ для критерия выбирается как

$$\pi(Z) = \begin{cases} 1, Z \in D_{t\alpha} \\ p, Z \in \Gamma_\alpha, \\ 0, Z \in C_{t\alpha} \end{cases}$$

а $\alpha = pF_0(\Gamma_\alpha) + F_0(D_{t\alpha})$, то критерий также является наиболее мощным.

В данной задаче под основной гипотезой понимается отсутствие встроенной информации (пустой контейнер) и альтернативной гипотезой – ее наличие (стего).

Теорема 3. Пусть в стегосистеме распределение стего является δ-приближенным относительно распределения контейнеров и выполняется критерий Неймана–Пирсона для заданного уровня значимости α. Тогда выполняется критерий стойкости

$$D(P_C \,//\, P_S) < \frac{1}{\ln a}\left[(t_\alpha - 1) + \delta\alpha\right]. \tag{21}$$

Доказательство. Введем обозначение $\delta_q = \dfrac{P_S(q) - P_C(q)}{P_C(q)}$, тогда $\dfrac{P_S(q)}{P_C(q)} = \delta_q + 1$.

Теперь относительную энтропию между распределением контейнеров и распределением стего можно записать в виде

$$D(P_C \,//\, P_S) = \sum_{q \in Q} P_C(q) \log(1 + \delta_q) \ .$$

Воспользуемся неравенством (7)

$$D(P_C \,//\, P_S) = \sum_{q \in Q} P_C(q) \log(1 + \delta_q) <$$

$$< \frac{1}{\ln a} \left[\sum_{C_{t_\alpha}} P_C(q)\delta_q + \sum_{\Gamma_\alpha} P_C(q)\delta_q + \sum_{D_{t_\alpha}} P_C(q)\delta_q \right]. \qquad (22)$$

Рассмотрим значения δ_q в каждой из трех областей отдельно:

– в области C_{t_α} выполняется неравенство $\dfrac{P_S(q)}{P_C(q)} < t_\alpha$, следовательно, в этой области $\delta_q < t_\alpha - 1$;

– в области Γ_{t_α} выполняется равенство $\dfrac{P_S(q)}{P_C(q)} = t_\alpha$, следовательно, в этой области $\delta_q = t_\alpha - 1$;

– в области D_{t_α} выполняется неравенство $\dfrac{P_S(q)}{P_C(q)} > t_\alpha$, но, поскольку распределение стего является δ-приближенным относительно распределения контейнеров, выполняется неравенство $\delta > \delta_q > t_\alpha - 1$.

Теперь можно усилить неравенство (22):

$$D(P_C \,//\, P_S) < \frac{1}{\ln a} \left[\sum_{C_{t_\alpha}} P_C(q)(t_\alpha - 1) + \sum_{\Gamma_\alpha} P_C(q)(t_\alpha - 1) + \sum_{D_{t_\alpha}} P_C(q)\delta \right]. \qquad (23)$$

Вынесем в (23) константы за знак суммирования и воспользуемся равенством $\sum_A P_C(q) = F_C(A),\ A \subset Q$:

$$D(P_C \,//\, P_S) < \frac{1}{\ln a}\left[(t_\alpha - 1)\sum_{C_{t_\alpha}} P_C(q) + (t_\alpha - 1)\sum_{\Gamma_\alpha} P_C(q) + \delta\sum_{D_{t_\alpha}} P_C(q)\right] =$$

$$= \frac{1}{\ln a}\left[(t_\alpha - 1)F_C(C_{t_\alpha} \cup \Gamma_\alpha) + \delta F_C(D_{t_\alpha})\right].$$

Поскольку при выполнении критерия Неймана–Пирсона $\alpha = pF_0(\Gamma_\alpha) + F_0(D_{t\alpha}) \geq F_0(D_{t\alpha})$, полученное неравенство можно усилить еще больше:

$$D(P_C \,//\, P_S) < \frac{1}{\ln a}\left[(t_\alpha - 1)F_C(\Gamma_\alpha \cup C_{t_\alpha}) + \delta\alpha\right]. \qquad (24)$$

Учитывая, что $F_C(\Gamma_\alpha \cup C_{t_\alpha}) \leq 1$, можно усилить и последнее неравенство, окончательно избавляясь от вероятностного распределения контейнеров:

$$D(P_C \,//\, P_S) < \frac{1}{\ln a}\left[(t_\alpha - 1) + \delta\alpha\right].$$

Теорема доказана.

Необходимо сравнить неравенства (21) с неравенством в теореме 1. В общем виде установить связь между значениями α и t_α мы не можем, но, очевидно, что уменьшение α ведет к росту t_α. Однако понятно, что порог t_α должен выбираться большим 1 и не превышать величины δ+1. Как правило, величина $\alpha \in [0,05; 0,01]$, т.е. $\alpha << 1$, и при $t_\alpha \approx 1$, правая часть неравенства (21) может оказаться намного меньшей, чем в неравенстве (2) теоремы 1.

Если при построении стегосистемы имеет значение оптимизация стойкости при пассивных атаках и уровня значимости критерия Неймана–Пирсона в задаче обнаружения встроенной информации, то с помощью неравенства (21) можно найти оптимальное решение.

Заключение

Для теоретико-информационной модели стегосистемы исследовался критерий стойкости при пассивных атаках, основанный на ограниченности относительной энтропии между распределением контейнеров и распределением стего. Было введено определение распределения, δ-приближенного относительно известного или желаемого распределения. Использование δ-приближенного распределения позволило получить простое решение задачи оценивания стойкости стегосистемы. Также для δ-приближенного распределения была установлена связь между параметрами критерия Неймана–Пирсона и критерием стойкости; получено неравенство, позволяющее оптимизировать стойкость стегосистемы и уровень значимости критерия Неймана–Пирсона.

В реальных условиях атаки на стегосистемы часто являются активными (масштабирование, сглаживание, преобразование сигналов для сжатия объемов передаваемой или хранимой информации и т. д.). Исследование стойкости стегосистем к различным видам активных атак будет представлено в дальнейших работах.

Л.Л. Нікітенко

ДОСЛІДЖЕННЯ КРИТЕРІЮ СТІЙКОСТІ ЩОДО ПАСИВНИХ АТАК

В роботі наведено визначення імовірнісного розподілу, δ-наближеного відносно відомого розподілу. Для δ-наближеного розподілу отримані оцінки стійкості стегосистеми для критерію стійкості щодо пасивних атак. Також встановлено зв'язок між параметрами критерію Неймана–Пірсона та критерієм стійкості щодо δ-наближеного розподілу. Отримано нерівність, що дозволяє оптимізувати стійкість стегосистеми та рівень значимості критерію Неймана–Пірсона.

L.L. Nikitenko

THE STABILITY CRITERION RESEARCH TO THE INACTIVE ATTACKS

The probability distribution is presented that is δ-approximated with regard to the well-known distribution. The stegosystem stability estimates of the stability criterion to the inactive attacks are derived while the distribution is δ-approximated. Furthermore, the characteristic of the stability criterion dependence of the Neumann–Pearson criterion characteristics is derived while the distribution is δ-approximated, the inequality is obtained that permits to optimize stegosystem stability and significance level of the Neumann–Pearson criterion.

1. *Грибунин В.Г.*, *Оков И.Н.*, *Туринцев И.В.* Цифровая стеганография. – М.: СОЛОН-Пресс, 2002. – 261 с.
2. *Задирака В.К.*, *Никитенко Л.Л.* К вопросу о стойкости стегосистем к обнаружению факта передачи скрываемых сообщений для двух частных случаев // Проблемы управления и информатики. – 2008. – № 3. – С. 152–156.
3. *Корн Г*,, *Корн Т.* Справочник по математике для научных работников и инженеров. – М., 1978. – 832 с.
4. *Гихман И.И.*, *Скороход А.В.*, *Ядренко М.И.* Теория вероятностей и математическая статистика. – Киев: Вища шк., 1979. – 408 с.

Получено 10.10.2008

Об авторе:

Никитенко Любовь Леонидовна,
младший научный сотрудник отдела оптимизации численных методов
Института кибернетики имени В.М. Глушкова НАН Украины.
e-mail zvk140@ukr.net