Лекция 7. Методы криптоанализа при неравновероятностной гамме. Расстояние единственности

Цель лекции: рассмотреть один из универсальных методов криптоанализа.

План лекции:
Введение.
1 Методы криптоанализа при неравновероятностной гамме. Расстояние единственности
Заключение
Контрольные вопросы

Ключевые слова: [выбрать самостоятельно].

Содержание лекции:

Введение

1 Методы криптоанализа при неравновероятностной гамме. Расстояние единственности

Пусть открытые тексты – слова длины n в алфавите Z/m. Для того, чтобы объяснить математическую теорию, разработанную для решения задачи, предположим, что гамма может принимать только r значений. Множество открытых X_n и шифрованных текстов Y_n погружены в пространство V_n всех последовательностей длины n в алфавите Z/m. Следуя [1, стр. 375], операцию дешифрования можно представить графически в виде ряда линий, идущих от каждого шифртекста $y \in Y_n$ к различным $x \in V_n$. В сделанных выше предположениях каждый $y \in Y_n$ связан ровно с $k = r^n$ различными $x \in V_n$. Обозначим их $x^{(1)}$, ..., $x^{(k)}$. Среди линий, идущих от y к $x^{(1)}$, ..., $x^{(k)}$ имеется открытый текст x_0. Если мы знаем признаки открытого текста (например, читаемость) и среди $x^{(1)}$, ..., $x^{(k)}$ ровно один текст x_0 удовлетворяет этим признакам, то тем самым процедура дешифрования завершена. Однако возможна ситуация, когда несколько текстов среди $x^{(1)}$, ..., $x^{(k)}$ удовлетворяют признакам открытого текста. Тогда цель дешифрования – получение достоверной информации из раскрытого открытого текста – не достигнута. Ясно, что на возможный исход дешифрования влияет r. Если r = 1, то от y ведет всего одна линия к x и, по условию, этот текст и есть открытый. Проблемы неоднозначности здесь нет. Если r = m, то от y ведет m^n различных линий ко всем элементам V_n. Значит, любой текст из X_n является возможным открытым текстом, что не дает нам возможности достигнуть цели дешифрования и этот случай обязательно порождает неоднозначность прочтения криптограммы. Для того, чтобы понять при каких r возможно однозначное дешифрование, а при каких r невозможно, рассмотрим следующую модель. На множестве ключей задана равномерная мера, т.е. любая допустимая гамма появляется с вероятностью 1/|K|. Мы также предположим, что для любого заданного шифртекста y, полученного из открытого текста x_0, все линии, связывающие y с $x^{(1)}$, ..., $x^{(k)}$ кроме (y, x_0), получены случайным и равновероятным выбором с возвращением из $(m^n - 1)$ возможностей. Это предположение было впервые введено Шенноном [2, стр. 374] при определении «случайного шифра» (фактически мы рассматриваем частный случай «случайного шифра»). Если модель с этим и следующими предположениями позволит оценивать r, допускающее однозначное дешифрование, и полученная оценка может быть подтверждена экспериментами, то теория правильно отражает суть вопроса.

---

[1] Шеннон К. Работы по теории информации и кибернетике, М.: Иностранная литература, 1963.
[2] там же

Обозначим ξ_i случайную величину, равную 1, если $x^{(i)}$ является открытым текстом, и 0 в противном случае. Тогда число открытых текстов η без x_0 среди $x^{(1)}, \ldots, x^{(k)}$ равно

$$\eta = \sum_{i=1}^{k} \xi_i - 1.$$

Тогда

$$E\eta = \sum_{i=1}^{k} E\xi_i - 1.$$

Для всех $x^{(i)}$, которые не равны x_0,

$$E\xi_i = \frac{|X|}{m^n - 1}.$$

В случае $x^{(i)} = x_0$, $E\xi_i = 1$. Тогда

$$E\eta = (k-1)\frac{|X|}{m^n - 1}. \quad (1)$$

Если соотношение параметров таково, что $E\eta << 1$, то по оценке Маркова

$$P(\eta \leq 1) \geq E\eta << 1.$$

Это означает, что появление ложных открытых текстов маловероятно или что x_0 выделяется однозначно с большой вероятностью. Для изучения соотношений между параметрами необходимо оценить |X|.

До сих пор мы представляли множество открытых сообщений x как подмножество множества V_n всех слов длины n. Однако нет оснований считать, что какая-либо последовательность не может стать сообщением. Просто вероятность встретить в качестве сообщения одни слова много больше, чем другие. Поэтому предполагают, что все последовательности длины n упорядочены в соответствии с их вероятностями появления в общении на данном языке. Тогда собственно осмысленные (в бытовом понимании этого слова) выражения относятся к более вероятным, чем последовательности, смысл которых нам неясен (трудно представить себе общение при помощи труднопроизносимых буквосочетаний).

Поскольку значения k, m^n и |X| велики, то оценку |X| можно сделать асимптотически, предполагая n большим. Такие оценки Шеннон разработал в работе «Математическая теория связи» [3, с. 265], к которой сейчас и обратимся.

Пусть

$$p(a_1), \ldots, p(a_m) \quad (2)$$

вероятности появления букв на фиксированном месте i в открытом сообщении длины n ($1 < i < n$). Предположим, что буквы в сообщении появляются независимо друг от друга с одним и тем же распределением (2). Подобная модель открытых сообщений является грубой, но, как и выше, оценка для r, близкая к экспериментальной, покажет, что наш

[3] Шеннон К. Работы по теории информации и кибернетике, М.: Иностранная литература, 1963.

механизм анализа отражает сущность проблемы. Обозначим через v_i, i = 1, ..., m, частоты букв a_1, …, a_m в последовательности x. Тогда вероятность выбора x в нашей схеме равна

$$P(x) = p^{v_1}(a_1) \dots p^{v_m}(a_m).$$

Будем считать, что

$$p(x) > 0, \quad H = -\sum_{i=1}^{m} p(a_i)\log_2 p(a_i).$$

Теперь мы можем сформулировать следующую теорему Шеннона.

Теорема 1 (К. Шеннон). Для любых ε >0 и δ > 0 можно найти такое n_0, что для любого n > n_0 последовательности из V_n распадаются на два непересекающихся класса B и $\overline{B}$ так, что

1) $P(\overline{B}) < \varepsilon$
2) $\forall x \in B$ выполняется неравенство

$$\left|\frac{\log_2 P^{-1}(x)}{n} - H\right| < \delta.$$

Доказательство. Возьмем произвольные малые ε > 0 и δ > 0 и рассмотрим события

$$\overline{B}_i = \{x \in V, |v_i - np(a_i)| > \delta n\}, \;\; i = 1, \dots, m.$$

По закону больших чисел $\exists n_0^{(i)}$, что $\forall n > n_0^{(i)}$

$$P(\overline{B}_i) < \frac{\varepsilon}{m}. \tag{3}$$

Определим $\overline{B} = \bigcup_{i=1}^{m} \overline{B}_i$. Тогда из (3) для $n > \max_i \left\{n_0^{(i)}\right\}$

$$P(\overline{B}) \le \sum_{i=1}^{m} P(\overline{B}_i) < \varepsilon. \tag{4}$$

Множество $B = \bigcap_{i=1}^{m} B_i$ может быть представлено в следующем виде:

$$B = \{x \in V, \forall i (i = 1, \dots, m) | v_i - np(a_i)| \le \delta n\}. \tag{5}$$

Обозначим $\alpha_i = v_i - np(a_i), \;\; i = 1, \dots, m$. Из (5) следует, что $\forall i \;\; -\delta n \le \alpha_i \le \delta n$. Выразим P(x), x ∈ B, через введенные параметры.

$$P(x) = \left(p(a_1)\right)^{np(a_1)+\alpha_1} \dots \left(p(a_m)\right)^{np(a_m)+\alpha_m}.$$

Тогда

$$\log_2 \frac{1}{P(x)} = -n \sum_{i=1}^{m} p(a_i) \log_2 p(a_i) - \sum_{i=1}^{m} \alpha_i \log_2 P(a_i).$$

Получаем, что $\forall x \in B$

$$\left|\frac{\log_2 P^{-1}(x)}{n} - H\right| \le \frac{1}{n}\sum_{i=1}^{m}|\alpha_i||\log_2 P(a_i)| \le \delta \sum_{i=1}^{m}|\log_2 p(a_i)|.$$

Поскольку $p(a_i) > 0$, то $\sum_{i=1}^{m}|\log_2 p(a_i)| = \text{const}$, не зависит от n. Теорема доказана.

Пусть $0 < \varepsilon < 1$, – произвольное малое число. Пусть все последовательности длины n расположены в порядке убывания вероятностей их появления. Как отмечалось выше, множество открытых сообщений моделируется начальным участком таких последовательностей. Обозначим через $\beta_n(\varepsilon)$ – число наиболее вероятных последовательностей таких, что сумма их вероятностей $\ge 1 - \varepsilon$, а сумма вероятностей любого набора из $(\beta_n(\varepsilon) - 1)$ этих последовательностей $< 1 - \varepsilon$. Следующая теорема показывает, что при $n \to \infty$ да множество последовательностей, составляющих в нашей модели открытый текст, не зависит от ε.

Теорема 2 (К. Шеннон). Для любого $\varepsilon > 0$

$$\lim_{n\to\infty}\frac{\log_2 \beta_n(\varepsilon)}{n} = H.$$

Докзательство. Пусть $Q_n(\varepsilon)$ - множество наиболее вероятных последовательностей таких, что $P(Q_n(\varepsilon)) \ge 1 - \varepsilon$, а исключение любой по-следовательности из $Q_n(\varepsilon)$ образует множество, вероятность которого меньше $1 - \varepsilon$. Возьмем малое $\delta > 0$ и рассмотрим множество B из предыдущей теоремы. Тогда для $x \in B$

$$H - \delta < \frac{-\log_2 P(x)}{n} < H + \delta$$

или

$$-n(H - \delta) > \log_2 P(x) > -n(H + \delta)$$

Отсюда

$$2^{-n(H+\delta)} < P(x) < 2^{-n(H-\delta)}. \qquad (6)$$

Рассмотрим оценки для $Q_n(\varepsilon)$. По определению $Q_n(\varepsilon)$ $P(Q_n(\varepsilon)) \ge 1 - \varepsilon$. Поскольку

$$Q_n(\varepsilon) = \left(Q_n(\varepsilon) \cap \overline{B}\right) \cup (Q_n(\varepsilon) \cap B),$$

тогда

$$P\left(Q_n(\varepsilon)\right) = \sum_{x\in Q_n(x)} P(x) = \sum_{x\in Q_n(x)\cap B} P(x) + \sum_{x\in Q_n(x)\cap \overline{B}} P(x).$$

Для второй суммы справедлива оценка

$$P\left(Q_n(\varepsilon) \cap \overline{B}\right) \le P\left(\overline{B}\right) < \varepsilon.$$

$$P(Q_n(\varepsilon) \cap B) = \sum_{x \in Q_n(x) \cap B} P(x) < \sum_{x \in Q_n(x) \cap B} 2^{-n(H-\delta)} =$$
$$= |Q_n(\varepsilon) \cap B| \cdot 2^{-n(H-\delta)} \leq |Q_n(\varepsilon)| \cdot 2^{-n(H-\delta)} =$$
$$= \beta_n(\varepsilon) 2^{-n(H-\delta)}$$

Следовательно,

$$1 - \varepsilon \leq P\big(Q_n(\varepsilon)\big) < \beta_n(\varepsilon) 2^{-n(H-\delta)} + \varepsilon.$$

Отсюда получаем

$$\beta_n(\varepsilon) > 2^{n(H-\delta)}(1 - 2\varepsilon).$$

Докажем, что множество $Q_n(\varepsilon)$ содержит минимальное число последовательностей среди всех множеств C, таких что $P(C) > 1 - \varepsilon$. Доказательство будем вести от противного. $\forall x \in Q_n(\varepsilon), x \notin C$ и $\forall y \in C, y \notin Q_n(\varepsilon)$, такие последовательности, что $P(x) \geq P(y)$, так как в $Q_n(\varepsilon)$ содержатся наиболее вероятные последовательности. Предположим, что

$$|C \backslash Q_n(\varepsilon)| < |Q_n(\varepsilon) \backslash C|$$

(т.е. $Q_n(\varepsilon)$ - не минимальное множество). Тогда

$$\sum_{x \in C \backslash Q_n(\varepsilon)} P(x) < \sum_{x \in Q_n(\varepsilon) \backslash C} P(x),$$

так как справа слагаемых хотя бы на 1 больше и все они не меньше слагаемых в левой сумме. Пусть y_0 имеет наименьшую вероятность среди всех $y \in Q_n(\varepsilon) \backslash C$. Следовательно,

$$\sum_{x \in C \backslash Q_n(\varepsilon)} P(x) \leq \sum_{x \in (Q_n(\varepsilon) \backslash C) \backslash \{y_0\}} P(x).$$

Поскольку

$$Q_n(\varepsilon) \backslash \{y_0\} = (Q_n(\varepsilon) \cap C) \cup \big((Q_n(\varepsilon) \backslash C) \backslash \{y_0\}\big),$$

то, учитывая, что по определению множества $Q_n(\varepsilon)$

$$P(Q_n(\varepsilon) \backslash \{y_0\}) < 1 - \varepsilon,$$

получим

$$P(C) = P(Q_n(\varepsilon) \cap C) + P\big(C \backslash Q_n(\varepsilon)\big) \leq P(Q_n(\varepsilon) \cap C) +$$
$$+ P(Q_n(\varepsilon) \backslash C \backslash \{y_0\}) = P(Q_n(\varepsilon) \backslash \{y_0\}).$$

Тогда

$$P(C) < 1 - \varepsilon,$$

что противоречит предположению. Следовательно, $Q_n(\varepsilon)$ – минимальное множество векторов, имеющих максимальную вероятность. Отсюда и из (6)

$$|Q_n(\varepsilon)| \le |B| = \sum_{x \in B} 1 < \sum_{x \in B} \frac{p(x)}{2^{-n(H+\delta)}} \le 2^{n(H+\delta)}.$$

Таким образом,

$$2^{n(H-\delta)} < |Q_n(\varepsilon)| < 2^{n(H+\delta)}$$

или

$$H - \delta < \frac{\log_2 \beta_n(\varepsilon)}{n} < H + \delta.$$

Следовательно,

$$\lim_{n \to \infty} \frac{\log_2 \beta_n(\varepsilon)}{n} = H,$$

что и требовалось доказать.

Множество открытых текстов можно представить как объединение B и $\overline{B}$. Множество $\overline{B}$ имеет маленькую вероятность, множество B оценивается по теореме 2.

$$|X| \approx 2^{nH(x)}.$$

Следовательно,

$$E\eta = (r^n - 1)\frac{|X|}{m^n - 1} \to 0$$

при

$$E\eta = \left(\frac{r}{m} \cdot 2^H\right)^n \to$$

Это выполняется, когда

$$\frac{r}{m} \cdot 2^H < 1,$$

то есть при r, удовлетворяющих неравенству

$$\log_2 r - \log_2 m + H < 0, \qquad (7)$$

возможно однозначное дешифрование.

Заключение

Контрольные вопросы

Смотри руководство по организации самостоятельной работы магистрантов.