ВЕСТНИК ПЕРМСКОГО УНИВЕРСИТЕТА
2010 Математика. Механика. Информатика Вып. 1

УДК 681.3.001.63; 004.056.55

Генетический алгоритм для криптоанализа шифра Виженера

В. В. Морозенко, Г. О. Елисеев
Пермский государственный университет, 614990, Пермь, ул. Букирева, 15

Разработан и описан генетический алгоритм для поиска секретного ключа шифра Виженера. Ключом является случайная последовательность символов из заданного алфавита, а исходными данными для криптоанализа – зашифрованный текст. С помощью разработанного генетического алгоритма задача криптоанализа решается в два этапа: на первом этапе вычисляется длина ключа, а на втором – сам ключ. Таким образом процесс расшифрования текста удается почти полностью автоматизировать.

Введение

Шифрование информации в настоящее время стало едва ли не основным методом её защиты. Доступность вычислительной техники и стремительный прогресс в её развитии привели к совершенствованию давно известных шифров и применению в массовом масштабе новых высоконадежных схем шифрования. Однако у этого прогресса есть и другая сторона: возросшие возможности вычислительных устройств сегодня успешно применяются не только для шифрования, но и для «взлома» тех шифров, которые ещё вчера, казалось бы, гарантировали полную защиту информации.

Исторически первый вариант шифра Виженера упоминается Юлием Цезарем в «Записках о галльской войне» (I в. до н.э.). Более сложная модификация этого шифра встречается в романе Ж.Верна «Жангада». Современные варианты шифра Виженера часто используются в коммерческих программных продуктах, например, в системах MS DOS и Macintosh.

При шифровании с помощью шифра Виженера используется секретный ключ – символьная строка. Чем больше её длина, тем надежней защищен зашифрованный текст от попытки его расшифровать (криптоатаки). Задача криптоанализа состоит в том, чтобы расшифровать зашифрованный текст, не обладая знанием секретного ключа. Решая эту задачу, злоумышленник, как правило, сначала находит секретный ключ, а затем с его помощью расшифровывает перехваченное зашифрованное сообщение.

Постановка задачи

Шифр Виженера – это блочный многоалфавитный подстановочный шифр [1]. Для удобства будем считать, что исходный текст представляет собой строку $(x_1, x_2, ..., x_m)$ длины m, образованную символами алфавита $A = \{a_1, a_2, ..., a_s\}$. Через $\nu(x)$ обозначим номер символа x в алфавите A и положим $\nu(a_i) = i - 1$. Секретным ключом K является строка $(k_1, k_2, ..., k_n)$ длины n, составленная из символов алфавита A.

Шифрование исходного текста с помощью данного ключа осуществляется следующим образом. Исходный текст разбивается на блоки длины n. Первый блок – это символьная строка $(x_1, x_2, ..., x_n)$, второй блок – строка $(x_{n+1}, x_{n+2}, ..., x_{2n})$ и т.д. Чтобы зашифровать первый блок, для каждого $i = 1, 2, ..., n$ символ x_i заменяем на символ

z_i такой, что номера $\nu(x_i), \nu(k_i), \nu(z_i)$ символов x_i, k_i и z_i в алфавите A связаны соотношением

$$\nu(z_i) \equiv (\nu(x_i) + \nu(k_i)) (\mathrm{mod}\ s). \qquad (1)$$

Тогда первый блок зашифрованного текста будет иметь вид $(z_1, z_2, ..., z_n)$. Аналогично происходит шифрование второго и всех последующих блоков исходного текста: символ z_j зашифрованного текста, где $j = 1, 2, ..., m$, однозначно определяется через свою позицию j в зашифрованном тексте и свой номер $\nu(z_j)$ в алфавите A. Правило для вычисления $\nu(z_j)$ обобщает формулу (1) и можен быть задано следующим образом:

$$\begin{cases} \nu(z_j) \equiv (\nu(x_j) + \nu(k_i)) (\mathrm{mod}\ s), \\ j \equiv i\ (\mathrm{mod}\ n), \\ j = 1,2,...,m, \\ i = 1,2,...,n. \end{cases} \qquad (2)$$

Для примера зашифруем с помощью шифра Виженера следующий текст:

«греция_становится_центром_ремесленной_индустрии_и_торговли._осуществляется_переход_от_первобытнообщинного».

Будем считать, что алфавит A состоит из 36 символов: 33 букв русского алфавита и символов «_» (пробел), «.» (точка), «,» (запятая). Выберем естественную нумерацию символов алфавита A: $\nu(а) = 0$, $\nu(б) = 1$, $\nu(в) = 2$, …, $\nu(э) = 30$, $\nu(ю) = 31$, $\nu(я) = 32$, $\nu(\_) = 33$, $\nu(.) = 34$, $\nu(,) = 35$. Если в качестве ключа взять слово «*математика*» и применить правило (2) с параметрами $s = 36$, $n = 10$, $m = 105$, то получим зашифрованный текст

«прчыхяпъэаъофнясоё.еътаущ_анчеюлч
тъоьёунрубчэиыёу_ырхуолыжзоюуи_йю
тффжеясовьеан_ор_.чйпчщмонывтыоу
,унъох».

Расшифровать текст, зашифрованный с помощью шифра Виженера, зная секретный ключ K, несложно. Для этого надо выполнить действия, обратные тем, которые применялись при его шифровании, т.е. для каждого $j = 1, 2, ..., m$ найти $\nu(x_j)$ по формуле

$$\nu(x_j) \equiv (\nu(z_j) - \nu(k_i)) (\mathrm{mod}\ s),$$

где $j \equiv i\ (\mathrm{mod}\ n)$.

Рассматриваемая в данной статье задача криптоанализа состоит в том, чтобы, имея зашифрованный текст $(z_1, z_2, ..., z_m)$, но не зная ни самого ключа K, ни его длины n, восстановить исходный текст $(x_1, x_2, ..., x_m)$. При этом предполагается, что злоумышленнику известен алфавит A, но не известна нумерация символов, использованная шифровальщиком. Простые расчеты, приведенные в [2] для простейшего алфавита $A = \{0,1\}$, показывают, что если секретный ключ представляет собой 56-битовую строку, то попытка подобрать его, наугад перебирая все возможные комбинации и проверяя за одну секунду 8000 случайно выбранных ключей, может занять более 500 лет, и есть только один шанс и 200000 угадать ключ в течение первых суток.

Описание генетического алгоритма

В настоящее время существует несколько способов «взлома» шифра Виженера. Как правило, сначала пытаются определить длину ключа, а затем и сам ключ. Например, для вычисления длины ключа можно применить тест Казиски [1]. Криптоанализ шифра Виженера использует известное свойство любого достаточно длинного осмысленного текста, состоящее в том, что в таком тексте каждая буква встречаются с почти предсказуемой среднестатистической частотой, заранее известной и характерной для любого достаточно длинного осмысленного текста. Например, в текстах на русском языке среднестатистическая частота встречаемости буквы «*а*» равна 0.052, а буквы «*м*» – 0.026.

Предлагаемый в данной статье генетический алгоритм для криптоанализа шифра Виженера тоже использует указанную устойчивость частотных характеристик осмысленных текстов. Кроме того он опирается на следующее свойство модулярной арифметики. Если две независимые дискретные случайные величины ξ и η принимают значения из множества $M = \{0, 1, 2, …, n – 1\}$, причем величина ξ принимает любое значение из ука-

занного множества с одинаковой вероятностью, равной $1/n$, а величина η для каждого $i = 0,1,2,\ldots,n-1$ принимает значение i с вероятностью p_i, то случайная величина

$$\xi + \eta \pmod{n}$$

принимает все значения из множества M с одинаковой вероятностью, равной $1/n$, т.е. имеет равномерное дискретное распределение на множестве M. Действительно, для каждого фиксированного $i = 0,1,2,\ldots,n-1$ вероятность $P(\xi + \eta \equiv i \pmod{n})$ того, что сумма $\xi + \eta$ по модулю n окажется равной i, в силу независимости ξ и η удовлетворяет равенствам

$$P(\xi + \eta \equiv i \pmod{n}) =$$

$$= \sum_{k=0}^{n-1} P(\xi = k, \eta \equiv i - k \pmod{n}) =$$

$$=$$

$$\frac{1}{n} \cdot \sum_{k=0}^{n-1} P(\eta \equiv i - k \pmod{n}) = \frac{1}{n} \cdot \sum_{k=0}^{n-1} p_k = \frac{1}{n}.$$

Отмеченный математический факт означает, что если каждый блок длины n исходного текста «сложить» по модулю s с ключом K той же длины, все символы которого случайным образом с равной вероятностью выбираются из алфавита A, то и в полученном тексте все символы будут встречаться с равной вероятностью. Иными словами, если на осмысленный текст, в котором каждый символ x_i встречается со своей частотой p_i, близкой к известному среднестатистическому значению, наложить «шум», в котором появление всех символов равновероятно, то в результате получится «шум». Именно это свойство модулярной арифметики позволяет шифру Виженера «прятать» в зашифрованном тексте, который по своим частотным характеристикам напоминает обычный «шум», секретную информацию, содержащуюся в исходном тексте.

Генетические алгоритмы хорошо зарекомендовали себя при решении задач комбинаторной оптимизации, в которых требуется найти экстремум функции, заданной на дискретном множестве допустимых решений [3]. Поскольку цель «взлома» шифра – найти среди множества потенциальных ключей единственный секретный ключ, то задачу криптоанализа вполне можно отнести к задачам комбинаторной оптимизации и попытаться применить для её решения генетический алгоритм. Такие попытки известны, и они оказались успешными [4,5].

Разрабатывая генетический алгоритм для криптоанализа шифра Виженера, необходимо выбрать способ кодирования допустимых решений (потенциальных ключей), определить правила скрещивания и мутации особей, стратегии их отбора, а также на множестве всех допустимых решений задать фитнесс-функцию, чья точка глобального максимума (или минимума) и будет являться искомым секретным ключом.

Применим следующий способ кодирования ключей. Хромосома, кодирующая произвольный ключ $(b_1, b_2, \ldots, b_n)$, представляет собой числовой вектор

$$(\nu(b_1), \nu(b_2), \ldots, \nu(b_n)),$$

где $\nu(b_i)$ – номер символа b_i в алфавите A. Напомним, что символы алфавита A пронумерованы числами $0,1,2,\ldots,n-1$. Такой способ кодирования ключей позволяет использовать классические варианты операторов скрещивания и мутации. Мутация хромосомы происходит с вероятностью p_0 и состоит в том, что два её произвольных гена меняются местами. Поскольку гены независимы, возникающая после мутации особь будет жизнеспособной. Вероятность p_0, количество особей, участвующих в скрещивании, и численность популяции являются настраиваемыми параметрами алгоритма.

Выбор особей для скрещивания будем осуществлять по «принципу рулетки», согласно которому вероятность того, что данная особь будет участвовать в скрещивании, прямо пропорциональна уровню её приспособленности [3]. Скрещивание двух особей будем выполнять с помощью одноточечного кроссовера. При этом положение точки разбиения хромосом выбирается случайным образом, после чего хромосомы обмениваются своими начальными отрезками.

Чтобы выбрать подходящую фитнесс-функцию, следует руководствоваться двумя соображениями. Во-первых, искомое решение должно совпадать с точкой её глобального экстремума. Во-вторых, значение фитнесс-функции для каждой конкретной особи должно отражать уровень её приспособленности.

Причем небольшим изменениям хромосомы должны соответствовать небольшие изменения фитнесс-функции. Если считать, что исходный текст, который был зашифрован шифром Виженера, являлся осмысленным, и частоты встречаемости символов в нем были примерно равны среднестатистическим значениям, то в качестве фитнесс-функции можно взять функцию $F(K)$ из [6]. Она вычисляется по формуле

$$F(K) = \sum_{ij} \left| D_{ij}^{K} - E_{ij} \right|,$$

где D_{ij}^{K} – частота встречаемости двухбуквенного слога «$b_i b_j$» (биграммы) в тексте, полученном из зашифрованного текста после его расшифрования с помощью предполагаемого ключа K, а E_{ij} – частота встречаемости этой же биграммы в некотором «среднестатистическом» осмысленном тексте. Среднестатистические частоты встречаемости всех биграмм заранее известны (например, они имеются в [1]) или могут быть вычислены на основании статистического анализа большого числа достаточно длинных осмысленных текстов. Очевидно, чем ближе предполагаемый ключ к искомому секретному ключу, тем больше расшифрованный с его помощью текст будет похож по своим частотным характеристикам на осмысленный текст и, следовательно, тем меньше будет соответствующее ему значение фитнесс-функции.

Заметим, что теоретический минимум предложенной фитнесс-функция равен нулю, однако она не обращается в нуль даже при расшифровании текста с помощью настоящего секретного ключа. По этой причине решено завершать работу генетического алгоритма, как только число поколений превысит заранее заданный порог, не дожидаясь достижения теоретического минимума фитнесс-функции.

Результаты работы генетического алгоритма

Разработанный авторами данной статьи генетический алгоритм решает задачу криптоанализа в два этапа. На первом этапе он находит длину секретного ключа, а на втором этапе – сам ключ.

Чтобы найти длину секретного ключа, сначала генетический алгоритм запускают в предположении, что длина секретного ключа равна некоторому l, и после завершения работы алгоритма запоминают «рекордное» (т.е. наименьшее) значение $c[l]$ фитнесс-функции за все время его работы. Затем параметр l увеличивают на 1, 2, 3 и т.д., каждый раз заново запуская генетический алгоритм, а после его завершения вычисляют «рекордное» значение $c[l + 1]$, $c[l + 2]$, $c[l + 3]$ и т.д. фитнесс-функции для ключей длины $(l + 1)$, $(l + 2)$, $(l + 3)$ и т.д. Покажем, что наименьшее среди найденных «рекордных» значений фитнесс-функции достигается именно при совпадении проверяемой длины и длины секретного ключа. Действительно, пусть n – длина секретного ключа. Если числа n и l совпадают, то можно ожидать, что в результате работы генетического алгоритма будет найден сам секретный ключ или близкий к нему ключ, а фитнесс-функция при использовании этого ключа примет относительно небольшое значение.

Пусть $K = (k_1, k_2, \ldots, k_n)$ – секретный ключ. Рассмотрим ключ

$$K' = (k_1, k_2, \ldots, k_n, k_1, k_2, \ldots, k_n)$$

длины $l = 2n$, который является конкатенацией двух секретных ключей K. Поскольку оба ключа K и K' превращают исходный текст в один и тот же зашифрованный текст, то и расшифрован он будет правильно обоими ключами. Это означает, что при поиске ключа длины $2n$ генетический алгоритм найдет решение, близкое к ключу K'. Этому ключу будет соответствовать относительно небольшое значение фитнесс-функции, поэтому найденное алгоритмом «рекордное» значение фитнесс-функции для ключей длины $2n$ также будет относительно небольшим. То же самое справедливо и для ключей с длинами, кратными числу n. Иными словами, если проверяемая длина ключа l кратна длине настоящего секретного ключа n, то на выходе алгоритма следует ожидать ключ с относительно небольшим значением фитнесс-функции. Если же проверяемая длина ключа l не кратна длине секретного ключа n, а сам секретный ключ является случайной символьной строкой, в которой на любой позиции с равной вероятностью может оказаться любой символ алфави-

та A, то «рекордное» значение фитнесс-функции окажется заметно больше.

Таким образом идея, используемая при нахождении длины ключа, заключается в следующем. Многократно запуская генетический алгоритм в предположении, что длина секретного ключа равна l, и каждый раз увеличивая этот параметр на единицу, получаем последовательность $c[l]$, $c[l + 1]$, $c[l + 2]$ и т.д. из «рекордных» значений фитнесс-функции. В этой последовательности обязательно будет присутствовать подпоследовательность, состоящая из элементов, чьи номера образуют арифметическую прогрессию $j, j + d, j + 2d$ и т.д., а сами элементы этой подпоследовательности $c[j]$, $c[j + d]$, $c[j + 2d]$ и т.д. будут заметно отличаться в меньшую сторону от остальных элементов последовательности. Тогда разность прогрессии d окажется в точности равной искомой длине n секретного ключа.

После того, как длина ключа будет найдена, снова запускаем генетический алгоритм. Полученное им решение, соответствующее «рекордному» значению фитнесс-функции, либо окажется искомым секретным ключом, либо будет близко к нему.

Выше был приведен пример исходного текста, который затем зашифровали с помощью ключа «*математика*». Генетический алгоритм был запущен для расшифровки полученного зашифрованного текста с предполагаемой длиной l секретного ключа, равной 4, 5, 6, …, 34, и вычислены соответствующие «рекордные» значения фитнесс-функции $c[4]$, $c[5]$, $c[6]$, …, $c[34]$. Оказалось, например, что

$c[8] = 159.46$,
$c[9] = 148.49$,
$c[10] = 42.59$ (абсолютный «рекорд»),
$c[11] = 148.46$,
$c[12] = 156.63$.

На рис. 1 хорошо видны «провалы», которые соответствуют «рекордным» значениям фитнесс-функции, полученным для предполагаемых длин l, равных 10, 20 и 30. Поскольку эти числа образуют арифметическую прогрессию с разностью 10, то можно сделать правильный вывод, что искомая длина секретного ключа равна 10.

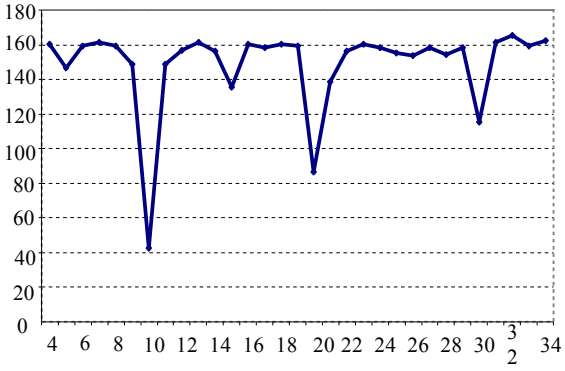

Рис. 1. *«Рекордные» значения фитнесс-функции для различных предполагаемых длин секретного ключа*

Повторная работа генетического алгоритма над тем же зашифрованным текстом, но с уже известной длиной секретного ключа позволила найти ключ «*математиёа*», который отличается от настоящего секретного ключа только в предпоследней позиции. В результате расшифрования с помощью этого ключа получен следующий текст:

«грецияˍсчановитсяˍыентромˍрессленноˍйˍнндустрииˍнˍорговли.восуществлбетсяˍˍпереъодˍотˍпержобытнообщннного».

Полностью расшифровать зашифрованный текст не удалось, поскольку полученный текст не совпадает с исходным. Однако можно заметить, что он очень близок к исходному. В нем можно заметить осмысленные («*греция*») или почти осмысленные слова («*ыентром*», «*ннндустрии*»). Благодаря наличию таких слов процесс расшифрования полученного текста можно довести до конца «вручную».

При работе алгоритма были выбраны следующие значения основных параметров: численность популяции – 70 особей, количество поколений – 40, вероятность мутации – 0.2, в каждом поколении 50% особей участвовали в скрещивании.

Отметим, что для ускорения работы алгоритма при поиске длины ключа можно использовать популяции с малой численностью и небольшое число поколений, а после нахождения истинной длины ключа n можно повторно применить генетический алгоритм с большими значениями указанных параметров для более точного отыскания секретного ключа.

Заключение

Проведенное исследование показало, что предложенный в данной статье генетический алгоритм вполне может быть использован для криптоанализа шифра Виженера, если верно предположение о том, что исходный текст, подвергшийся шифрованию, является осмысленным текстом достаточной длины и обладает среднестатистическим частотным профилем. Такое предположение вполне естественно и почти не сужает область применимости данного алгоритма. Более того, устойчивыми частотными характеристиками обладают все «живые» языки, поэтому алгоритм можно легко адаптировать для расшифрования текстов, написанных, например, на английском или французском языке. Для этого достаточно ввести в алгоритм информацию о среднестатистических частотах биграмм указанных языков.

Предложенный генетический алгоритм далеко не всегда полностью расшифровывает зашифрованное сообщение. Получаемый на выходе алгоритма текст чаще всего отличается от исходного текста, хотя и незначительно. Как правило, расшифрованный текст имеет неточности, которые не бросаются в глаза. При удачном подборе параметров генетический алгоритм показывает вполне хорошие результаты как по скорости сходимости, так и по качеству получаемого решения. Нельзя сказать, что применение описанного генетического алгоритма позволяет полностью автоматизировать процедуру криптоанализа, однако он сводит к минимуму участие человека и его «ручную» работу в этом процессе.

Частным случаем шифра Виженера является шифр Вернама, в котором длина секретного ключа совпадает с длиной исходного текста. К.Шеннон доказал, что шифр Вернама абсолютно надёжен [7]. Естественно, что и попытки «взломать» этот шифр с помощью описанного генетического алгоритма окажутся неудачными. Этот факт объясняется тем, что для успешной работы алгоритма необходимо, чтобы длина зашифрованного текста многократно превосходила длину секретного ключа. В ситуации с шифром Вернама это условие, очевидно, не выполняется.

Список литературы

1. *Алферов А.П., Зубов А.Ю., Кузьмин А.С., Черемушкин А.В.* Основы криптографии: Учебное пособие.– М.: Гелиос АРВ, 2001.
2. *Шнайер Б.* Прикладная криптография. Протоколы, алгоритмы, исходные тексты на языке Си.– М.: Издательство ТРИУМФ, 2002.
3. *Гладков Л.А., Курейчик В.В., Курейчик В.М.* Генетические алгоритмы / Под ред. В.М. Курейчика. – М.:ФИЗМАТЛИТ, 2006.
4. *Delman B.* GeneticAlgorithms in Cryptography // A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Engineering. New York, 2004.
5. *Городилов А.Ю.* Криптоанализ перестановочного шифра с помощью генетического алгоритма // Вестник ПГУ, Пермь, 2007, Вып. 7(12), с.44–49.
6. *Jakobsen T.* A Fast Method for the Cryptanalysis of Substitution Ciphers, 1995.
7. *Шеннон К.* Теория связи в секретных системах // В кн.: Работы по теории информации и кибернетике. – М.: ИЛ, 1963.

A genetic algorithm for cryptoanalysis of Vigener's cipher

V. V. Morozenko, G. O. Eliseev
Perm State University, 614990, Perm, Bukireva st., 15

A genetic algorithm for finding a secret key of Vigener's cipher is investigated and described. The key is a random sequence of symbols from a given alphabet, and input of the cryptanalysis is a ciphertext. The cryptanalysis problem is solved by the genetic algorithm in two stages: at the first stage the key length will be evaluated, and at the second stage – the key will be obtained. A given ciphertext will be decrypted by this automated process almost completely.