Detecção de padrões em criptogramas como suporte às atividades de criptoanálise

Renato Hidaka Torres
Instituto Militar de Engenharia
renatohidaka@gmail.com

Gláucio Alves de Oliveira
Instituto Militar de Engenharia
glaucioaorj@gmail.com

José A. M. Xexéo
Instituto Militar de Engenharia
Faculdade Salesiana Maria Auxiliadora
josexexeo@gmail.com

William A. R. Souza
Universidade Federal do Rio de Janeiro
william@cos.ufrj.br

Ricardo Liden
Faculdade Salesiana Maria Auxiliadora
rliden@pobox.com

Resumo

Várias pesquisas e ensaios em modernos processos de criptografia evidenciam indícios de assinatura associada ao tipo de algoritmo ou à chave utilizados no processo de cifragem. Em um ataque somente com texto cifrado, em que o criptoanalista dispõe da menor quantidade de informações, é necessário conhecer pelo menos o algoritmo criptográfico utilizado na cifragem. Esses indícios de identificação de padrões em criptogramas permitem ao criptoanalista duas possibilidades: a) Viabilizar um ataque somente com texto cifrado, por meio da integração dos ataques de distinção, ou identificação com os métodos tradicionais de criptoanálise; b) Contribuir para os critérios e testes de certificação estabelecidos pelo NIST (National Institute of Standards and Technology). Nesse contexto, este artigo apresenta os recentes avanços dos trabalhos nessa área. Em particular os desenvolvidos pelo Grupo de Segurança da Informação do Instituto Militar de Engenharia (GSI/IME), que corroboram a hipótese de existência de padrões em algoritmos criptográficos certificados pelo NIST.

Abstract

Several research and tests in modern encryption processes show evidence of signature associated with the type of algorithm or the key used in the encryption process. In a cipher text-only attack where the cryptanalyst have the least amount of information, it is necessary at least know the cryptographic algorithm used in encryption. This evidence of identified patterns allow the cryptanalyst two possibilities: a) Conduct an attack with only ciphertext, by integrating the attacks of distinction, or identification with traditional methods of cryptanalysis, b) Contribute to the criteria and certification tests provided by NIST (National Institute of Standards and Technology). Therefore, this article presents recent advances in research in this area. In particular, those developed by the Information Security Group of the Military Engineering Institute (GSI / EMI), which support the hypothesis of the existence of patterns in cryptographic algorithms certified by NIST.

1. Introdução

O desenvolvimento da criptografia foi estimulado pelo risco de interceptação de mensagens por pessoas não autorizadas e seu uso é conhecido a mais de 4 mil anos. A partir da década de 1970, o NBS (National Bureau of Standards) estabeleceu o primeiro algoritmo criptográfico padrão com base em cifras simétricas de blocos, o DES, o qual se destinava a proteger as comunicações de órgãos do governo e empresas privadas. Após mais de 20 anos de uso do DES, o NIST (National Institute of Standard and Technology) organizou um concurso para escolher o novo padrão criptográfico, também baseado em cifras simétricas de blocos, denominado AES (Advanced Encryption Standard), do qual saiu vencedor, em 2001, o algoritmo Rijndael. Nesse mesmo período, desenvolveu-se a criptografia assimétrica ou de chave pública, da qual o principal algoritmo é o RSA.

A identificação do algoritmo ou da chave criptográfica utilizada no processo de cifragem partindo do conhecimento apenas dos criptogramas por eles gerados é uma aspiração dos criptoanalistas e uma preocupação para os projetistas desse tipo de algoritmo. O tema é pouco abordado na literatura da área; mas, ultimamente, a aplicação de técnicas de Reconhecimento de Padrões nos criptogramas, evidenciou indícios de assinaturas decorrentes dos processos de cifragem, onde tais assinaturas são consequência dos algoritmos criptográficos ou das chaves.

Neste contexto, diversos pesquisadores detectaram padrões em criptogramas. (KNUDSEN, 2000) realiza o chamado ataque de distinção, para criptogramas gerados pelo algoritmo RC6 com o auxílio da estatística do qui-quadrado. Para este ataque, em particular, (UEDA, 2007) propôs um método para fortalecer o algoritmo RC6, como uma contramedida ao ataque com a estatística do qui-quadrado. (CARVALHO, 2006), (SOUZA, 2007) e (SOUZA, 2008) agruparam criptogramas gerados pelas cifras de blocos DES, AES e RSA, em função da chave, com diversos tamanhos de criptogramas e chaves. Outras técnicas foram usadas para identificar cifras, no modo ECB, com máquinas de vetor de suporte aplicadas aos

algoritmos DES, Triple DES, Blowfish, AES e RC5 (DILEEP, 2006); com testes de aleatoriedade, operações XOR e funções limiar, aplicadas aos algoritmos DES, IDEA, Blowfish, RC4, Camellia e RSA (MAHESHWARI, 2001), (CHANDRA, 2002), (RAO, 2003) e (SAXENA, 2008); e com métodos de histograma e de predição de bloco aplicados aos algoritmos DES, Triple DES, Blowfish, RC5 e AES (NAGIREDDY, 2008). (TORRES et al, 2010) desenvolveu algoritmos de agrupamento utilizando Algoritmo Genético e Teoria dos Grafos para agrupar as cifras geradas pelos cinco finalistas do concurso do AES com menor custo computacional. Em (SOUZA, 2010) foi proposto um mecanismo capaz de identificar *n*-cifras de blocos. O mecanismo foi aplicado em um caso particular de $n = 5$, demonstrando que pode ser feita a identificação de *n* cifras, onde *n* é um inteiro qualquer, a partir de um conjunto de criptogramas. Os resultados desses trabalhos reforçaram a hipótese da existência de assinaturas nos criptogramas decorrentes do processo de cifragem. Métodos de Inteligência Computacional também podem ser utilizados para verificação inicial da segurança de criptossistemas, buscando revelar padrões em seus criptogramas (LASKARI, 2007). Os trabalhos de (ALBASSAL, 2004), (DILEEP, 2008) e (RAO, 2009), por exemplo, apresentam o uso de métodos de Inteligência Computacional na criptoanálise de cifras de blocos baseadas na estrutura de Feistel.

Neste artigo será apresentada uma descrição dessas pesquisas, sendo detalhada a pesquisa realizada pelo GSI/IME.

2. Comentários dos trabalhos desenvolvidos pelo GSI/IME

Os trabalhos desenvolvidos pelo GSI/IME contribuíram para o reconhecimento de padrões em criptogramas e para a identificação de algoritmos criptográficos, por meio de procedimentos análogos aos ataques cuja informação é apenas de textos cifrados. (CARVALHO, 2006) foi fundamental desse processo, dando início a essa área de pesquisa no GSI/IME e contribuindo para a consolidação da fase de pré-processamento dos dados, haja vista que todos os trabalhos posteriores utilizaram a mesma modelagem para os criptogramas. (SOUZA, 2007) contribuiu através de uma profunda análise de medidas de similaridade e distância, como: Co-seno, Dice, Jaccard, Overlap, Simple-matching, distância Euclidiana, distância Manhattan, distância Canberra e distância Bray-Curtis; e de métodos de agrupamentos hierárquicos aglomerativos, como: single-link, complete-link e group-average link, confirmando o melhor desempenho da medida de similaridade co-seno e do método single-link dentro do contexto proposto. Além disso, (SOUZA, 2007) contribuiu por sua aplicação de redes neurais nos processos de agrupamento e classificação de criptogramas em função do algoritmo criptográfico. Essa foi a primeira tentativa do GSI/IME para a identificação de algoritmos criptográficos a partir somente do criptograma. (TORRES, 2010) é o trabalho mais recente desenvolvido pelo GSI/IME. A principal contribuição desse trabalho está na automatização do número de grupos a serem encontrados no processo de agrupamento e melhoria de desempenho referente ao tempo computacional. Diferentemente dos trabalhos anteriores, os algoritmos de agrupamento desenvolvidos por (TORRES, 2010) são algoritmo particionais. As técnicas utilizadas para a modelagem destes algoritmos foram Teoria dos Grafos e Algoritmo Genético.

Quanto às contribuições referentes aos experimentos, (CARVALHO, 2006), por ser o pioneiro, concluiu que o conjunto mínimo de amostras que poderia ser trabalhado era um conjunto de 1500 criptogramas de 512 bytes cada, totalizando uma amostra de 768000 bytes. (SOUZA, 2007) realizou muito mais experimentos e montou um banco de dados de criptogramas para teste. Além disso, conseguiu trabalhar com amostras menores do que (CARVALHO, 2006). (SOUZA, 2007) trabalhou com um conjunto de 150 criptogramas de 4 Kbytes cada, totalizando uma amostra de 614400 bytes. (TORRES et al, 2010) apesar de não conseguir trabalhar com amostras menores do que as já trabalhadas por (SOUZA, 2007), conseguiu com a mesma configuração obter agrupamentos de maior qualidade.

3. Contribuições do processo

A Figura 1 ilustra os componentes desenvolvidos pelo GSI/IME. A seguir será comentada a contribuição de cada autor na construção desses componentes.

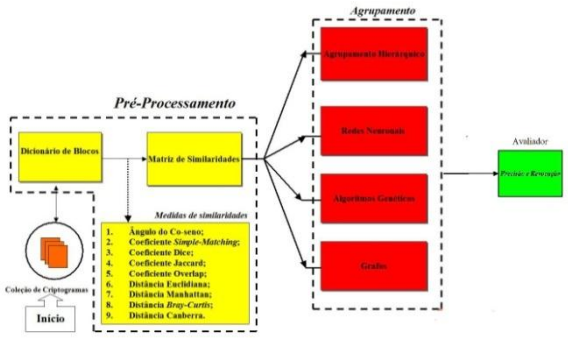


Figura 1. Componentes do processo

3.1 Modelagem do pré-processamento

Como já mencionado, a principal contribuição do trabalho de (CARVALHO, 2006) foi a modelagem da fase de pré-processamento. Esta fase é necessária para modelar os criptogramas em uma estrutura de dados passível de ser utilizada como dado de entrada por algoritmos de agrupamento.

O modelo considera uma coleção de criptogramas como um espaço de vetores de *n*-dimensões, onde *n* é o número de blocos binários no universo dos criptogramas da coleção. Deste modo, sejam dois criptogramas 1 e 2 em que a frequência $f_{n,1}$ é relacionada ao *n*-ésimo

bloco do criptograma 1 assim como a freqüência $f_{n,2}$ é relacionada ao *n*-ésimo bloco do criptograma 2 . Então, o vetor para o criptograma 2 é definido como $\vec{C}_2 = (f_{1,2}, f_{2,2}, ..., f_{n,2})$ e, da mesma forma, o vetor para o criptograma 1 é representado como $\vec{C}_1 = (f_{1,1}, f_{2,1}, ..., f_{n,1})$. A Figura 2 ilustra o processo de construção dos vetores dos criptogramas.

Em seguida, constrói-se um "dicionário" com *n* blocos binários, gerados pelo processo de contagem dos próprios blocos. O tamanho de cada bloco pode ser determinado por qualquer divisor do tamanho da chave. Após esta modelagem vetorial, foi implementada uma medida de associação entre os vetores, representada pelo co-seno entre eles (ver Equação 1). Após o cálculo de similaridade, é obtida uma matriz de similaridade simétrica *M* em que cada célula $M_{i,j}$ representa a similaridade entre o criptograma *i* e *j*. A Figura 3 ilustra a fase de pré-processamento.

Blocos binários	Criptogramas					
	$\vec{C}_1$	$\vec{C}_2$	$\vec{C}_3$	$\vec{C}_4$	.....	$\vec{C}_i$
11101001	$f_{1,1}$	$f_{1,2}$	$f_{1,3}$	$f_{1,4}$	.....	$f_{1,i}$
10101010	$f_{2,1}$	$f_{2,2}$	$f_{2,3}$	$f_{2,4}$	.....	$f_{2,i}$
:	:	:	:	:	.....	:
:	:	:	:	:	.....	:
:	:	:	:	:	.....	:
11100010	$f_{n,1}$	$f_{n,2}$	$f_{n,3}$	$f_{n,4}$	.....	$f_{n,i}$

Figura 2. Dicionário de blocos. $f_{n,1}$ é a freqüência do *n*-ésimo bloco do *i*-ésimo criptograma da coleção.

$$\text{Cos}(\vec{C}_1, \vec{C}_2) = \frac{\sum_{i=0}^{n}(\vec{C}_1 * \vec{C}_2)}{\sqrt{\sum_{i=0}^{n}\vec{C}_1^2 * \sum_{i=0}^{n}\vec{C}_2^2}} \qquad \text{(Equação 1)}$$

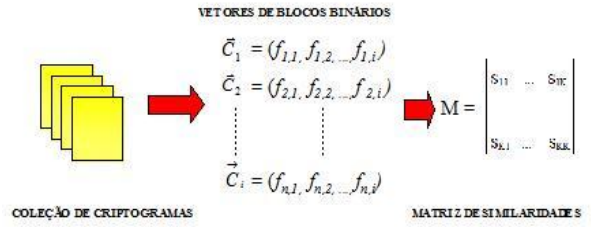
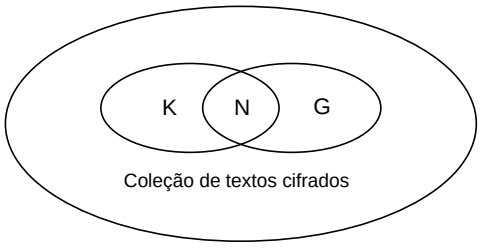


Figura 3. Fase de Pré-processamento.

3.2. Medidas de avaliação dos grupos

Na avaliação da qualidade do agrupamento foram utilizadas as medidas revocação e precisão (YATES e NETO, 1999) e (FUNG et al, 2003) (vide Figura 3.1).



K – Textos cifrados relevantes
G – Textos cifrados recuperados
N – Textos cifrados relevantes recuperados

Revocação = N/K
Precisão = N/G

Figura 3.1. Fase de Pré-processamento

Revocação indica a capacidade do método de recuperar todos os criptogramas relevantes. Precisão, por sua vez, indica a capacidade do método de recuperar apenas criptogramas relevantes.

Criptogramas relevantes são aqueles que pertencem naturalmente a um determinado grupo. Por exemplo, em um grupo formado por criptogramas cifrados pelo MARS, os criptogramas relevantes são os cifrados pelo MARS. Os demais criptogramas, cifrados por outros algoritmos, não são relevantes.

3.3. Agrupamento e classificação de chaves usando Rede Neural

A partir dos fundamentos gerados pelo trabalho de (CARVALHO, 2006), foi possível então, no trabalho de (SOUZA, 2007) testar medidas de similaridade e distâncias, assim como técnicas de agrupamento, as quais poderiam oferecer melhores resultados ao processo de agrupamento no contexto dos criptogramas. Como já mencionado, o melhor desempenho da medida de similaridade co-seno e do método single-link foi confirmado.

O próximo passo no processo seria a classificação dos criptogramas a partir dos grupos de chaves. O objetivo da classificação era que uma vez formados os grupos de chaves, um novo criptograma poderia ser alocado a um grupo já formado, indicando que tal criptograma foi cifrado pela mesma chave daquele grupo, construindo-se um dicionário de chaves. Desta forma, foi construída uma Rede Neural auto-organizável, baseada no mapa de Kohonen com a finalidade de agrupar e classificar as chaves (vide Figura 4).

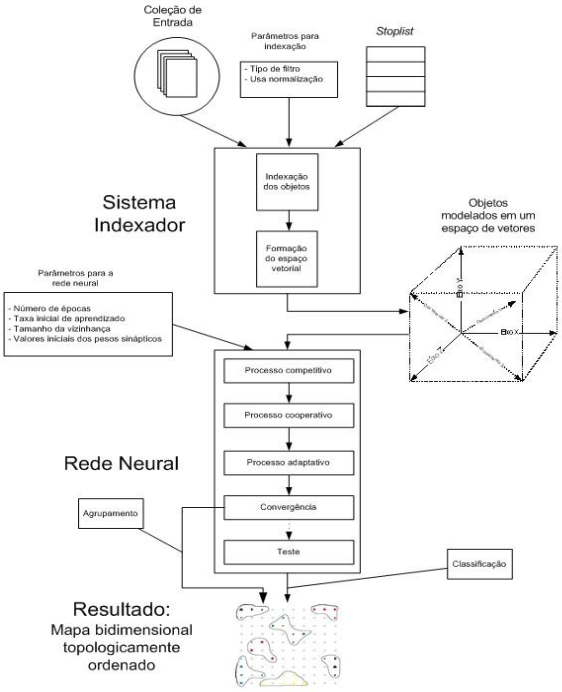


Figura 4. Fase de descrição da Rede Neural artificial desenvolvida para os experimentos.

Os resultados se mostraram promissores. Entretanto, devido a necessidade do elevado poder computacional para a execução da Rede Neural, não foi possível realizar experimentos que comprovassem o processo de classificação das chaves, dada que a amostra utilizada foi pequena. Como o objetivo final do processo de criptoanálise é obter o texto claro ou a chave utilizada na cifragem, o próximo passo seria a buscar maneiras de se identificar os algoritmos criptográficos. Desta forma, a partir de criptogramas gerados pelos algoritmos DES, AES e RSA, foram realizados experimentos de agrupamento e classificação de criptogramas em função do algoritmo criptográfico, utilizando a Rede Neural mencionada. Tais experimentos indicaram a possibilidade de identificação de algoritmos criptográficos a partir somente dos criptogramas (vide Figura 5).

Na Figura 5, as representações no mapa são feitas por meio de cores, onde os círculos em cinza-claro representam os criptogramas cifrados com o AES, os círculos em preto representando os criptogramas cifrados com o DES e os círculos em cinza representando os criptogramas cifrados com o RSA. Ainda neste trabalho foi construída a ferramenta WARSText (vide Figura 6), a qual automatizou os processo de agrupamento e classificação implementando as medidas de similaridades e distâncias: Co-seno, Dice, Jaccard, Overlap, Simple-matching, distância Euclidiana , distância Manhattan , distância Canberra e distância Bray-Curtis; os métodos de agrupamentos hierárquicos aglomerativos: single-link, complete-link e group-average link; e as técnicas de Rede Neural e de classificação Fuzzy por relações equivalentes.

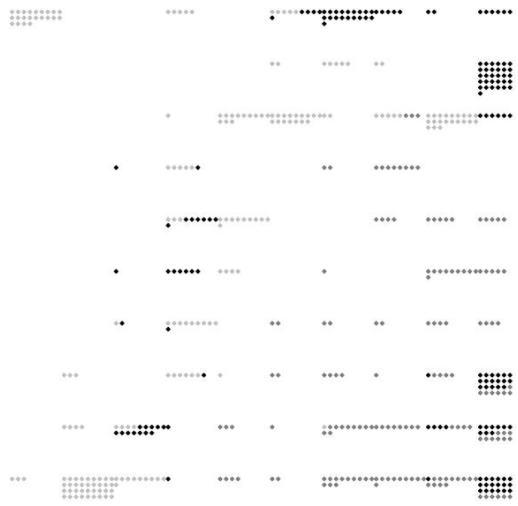


Figura 5. Mapa bidimensional com o agrupamento dos algoritmos AES, DES e RSA.

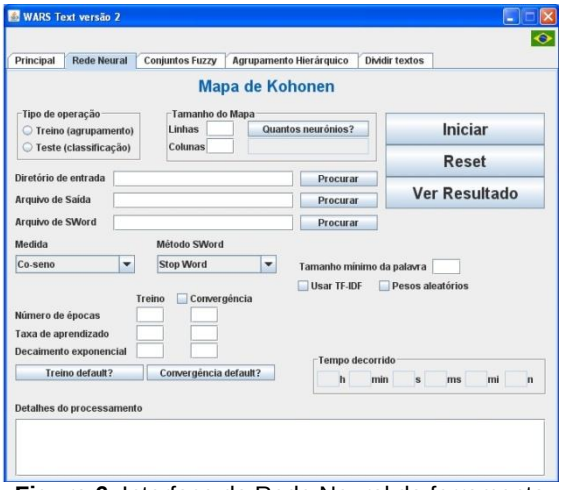


Figura 6. Interface da Rede Neural da ferramenta WARSText.

3.4. Agrupamento de cifras utilizando Algoritmo Genético e Grafos

(TORRES et al, 2010) como já mencionado, contribuíram, principalmente, com o agrupamento automatizado de criptogramas gerados por algoritmos ou chaves distintas. Diferentemente dos trabalhos anteriores, este trabalho utilizou algoritmos de agrupamento particional. As técnicas utilizadas para a construção dos algoritmos foram Algoritmo Genético (AG) e teoria dos Grafos.

3.4.1. Modelagem do Algoritmo Genético (AG)

O AG utiliza como entrada uma matriz de similaridades de um conjunto de criptogramas que está sendo analisado. Esta matriz é gerada pela fase de pré-processamento anteriormente vista na Figura 3.

a) Representação Cromossomial

Cromossomos são possíveis soluções do problema em questão e cada um deles representa um conjunto de grupos formado por criptogramas. Os cromossomos do AG modelado são gerados aleatoriamente de acordo com o modelo binário de (GOLDSCHMIDT, 2005). A Figura 7 ilustra o modelo cromossomial implementado. Cada linha da matriz representa um grupo e cada coluna um criptograma. Se um criptograma pertence a um determinado grupo, então o elemento da matriz que está no cruzamento da sua coluna com a linha do grupo terá valor igual a "1". Caso contrário, o elemento será igual a zero. Como cada criptograma pertence a um único grupo, cada coluna da matriz tem um elemento com valor "1", tendo o restante valor zero.

Grupos	Criptogramas				
	C_1	C_2	C_3	…..	C_i
Grupo 1	1	0	1	…..	0
Grupo 2	0	0	0	…..	0
Grupo 3	0	0	0	…..	0
Grupo 4	0	1	0	…..	0
:	:	:	:	:	:
Grupo k	0	0	0	…..	0

Figura 7. Modelo representativo do cromossomo do Algoritmo Genético.

b) Operadores Genéticos

- Crossover

O Crossover é o cruzamento genético de dois cromossomos (pais) que acarreta a geração de dois "novos cromossomos" (filhos). A taxa do operador de crossover foi arbitrada em 95%. Isto significa que a probabilidade de ocorrência de um cruzamento genético entre dois cromossomos é alta. Além da taxa do crossover, para a realização do cruzamento genético, é necessário determinar o "ponto de corte". Neste caso, o ponto é um intervalo entre dois genes de um cromossomo. Cada gene de um cromossomo representa um único criptograma. No AG modelado foi convencionado o esquema de crossover de "dois pontos de corte" aleatórios. Isto significa que durante as operações de crossover, os criptogramas dinamicamente mudarão de grupos. Nas Figuras 8 e 9, por exemplo, podemos observar que os criptogramas C_2 e C_3 mudam de grupo no cruzamento genético de dois cromossomos. Assim, após o crossover, temos dois "novos cromossomos" ou dois novos conjuntos de grupos de criptogramas.

- Mutação

O operador de mutação por definição atua sobre um determinado gene, alterando-lhe o valor de forma

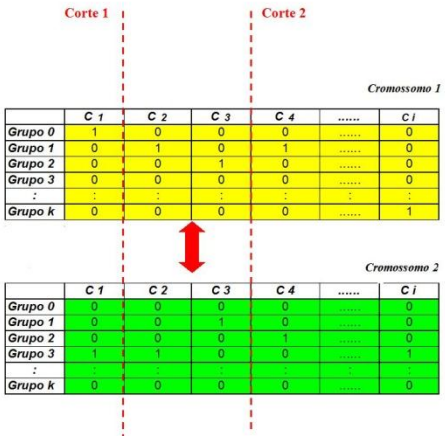


Figura 8. "**Dois pontos de corte" aleatórios. Os cromossomos 1 e 2 submetidos ao cruzamento genético formarão dois filhos**.

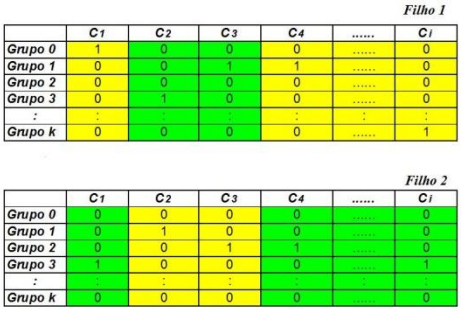
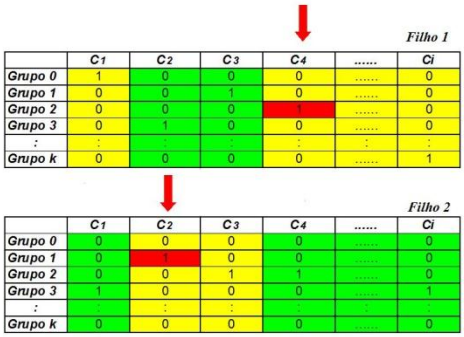


Figura 9. O filho 1 é formado pelo material genético do *cromossomo* 2 que está entre os "pontos de corte" mais o material genético do *cromossomo* 1 fora dos pontos de corte.

aleatória. Isto significa que após a operação de crossover, um criptograma é escolhido aleatoriamente, em cada novo cromossomo gerado, para mudar de grupo. Deste modo, no AG modelado para cada cromossomo criado após o crossover, um criptograma é escolhido aleatoriamente e os valores da sua correspondente coluna são zerados. Apos isto, é sobreposto o valor "1" em uma posição aleatória da coluna. Relembra-se que cada criptograma somente pode pertencer a um único grupo. Assim, existe a possibilidade de ocorrer uma inserção do valor "1" em uma determinada posição da coluna que já possuía tal valor. Neste caso, o cromossomo não sofre efetivamente uma mutação. Isto significa que determinado criptograma não mudou de grupo. Para ilustrar esta situação especifica, observamos nas Figuras 9 e 10 em que o criptograma C_2 (Filho 2) não foi classificado em outro grupo.

A taxa de mutação adotada no Algoritmo Genético é de 1%. Isto significa que a probabilidade de ocorrer a mudança de grupo de um determinado criptograma é pequena em relação à taxa de crossover. Observa-se na Figura 9 que o criptograma C_4 (Filho 1) que antes

pertencia ao Grupo 1, agora pertence ao Grupo 2 (vide Figura 10).



	C1	C2	C3	C4	.....	Ci
Grupo 0	1	0	0	0	.....	0
Grupo 1	0	0	1	0	.....	0
Grupo 2	0	0	0	1	.....	0
Grupo 3	0	1	0	0	.....	0
:	:	:	:	:	:	:
Grupo k	0	0	0	0	.....	1

	C1	C2	C3	C4	.....	Ci
Grupo 0	0	0	0	0	.....	0
Grupo 1	0	1	0	0	.....	0
Grupo 2	0	0	1	1	.....	0
Grupo 3	1	0	0	0	.....	1
:	:	:	:	:	:	:
Grupo k	0	0	0	0	.....	0

Figura 10. Operação de mutação após a operação de *crossover*.

c) Função de avaliação

O índice Calinski-Harabasz (CH) criado por (CALINSKI, 1974) é utilizado como função de avaliação. Esta função permite encontrar automaticamente o número correto de grupos homogêneos em um conjunto de criptogramas gerados por algoritmos criptográficos distintos ou pelo mesmo algoritmo criptográfico com chaves distintas. Assim, o índice CH faz a avaliação dos cromossomos gerados em que o mais bem avaliado representa o particionamento correto de todos os criptogramas que são mais similares. A equação (2) ilustra a função CH.

$$\frac{B(k-1)}{W(n-k)} = \frac{\sum_{i=1}^{k} n_i \|z_i - z\|^2 (k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i} \|x_{ij} - z_i\|^2 (n-k)} \qquad \text{(Equação 2)}$$

3.4.2. Modelagem do Grafo

Dado um Grafo não direcionado G(V,E) em que V = $x_1,x_2,...,x_n$ são os vértices e as arestas (x_i, x_j), cujo peso determina a similaridade entre os vértices aos quais ela está conectada, o objetivo é encontrar o número de subgrafos conexos pertencentes á *G*. A estrutura de dados utilizada para a representação do Grafo foi a matriz de similaridade obtida na fase de pré-processamento.

O pseudocódigo do algoritmo possui os seguintes passos:

1. Discretizar a matriz de similaridade.

2. Construir a lista de adjacência para cada vértice

3. Escolher uma lista e aplicar um busca em largura para encontrar todos os vértices conexos ao vértice cabeça da lista e marcar este vértice como já visitado.

4. Para cada vértice encontrado, repetir o passo 3 até que todos os vértices conexos sejam visitados.

5. Se todos os vértices já foram visitados então fim, senão voltar ao passo 3 para encontrar uma nova participação.

Após a execução do algoritmo, pode-se perceber que o número de grupos encontrados será igual ao número de criptogramas cifrados pela mesma chave ou algoritmo. A complexidade do algoritmo é $O(n^2)$, ficando esta complexidade em função da iteração dos vértices e da busca em largura para cada vértice ainda não visitado.

Esta modelagem é interessante devido a sua simplicidade e desempenho computacional, pois o algoritmo proposto possui o menor custo computacional dentre todos os algoritmos propostos pelo GSI/IME, além de apresentar resultados tão satisfatórios quanto os outros modelos no que diz respeito à qualidade dos grupos encontrados.

3.4.3. Construção de componentes

Outra contribuição resultante do trabalho de (TORRES et al, 2010) foi o desenvolvimento dos componentes visuais. O primeiro componente desenvolvido teve como objetivo permitir a visualização dos grupos e facilitar a análise dos mesmos, dando a possibilidade de verificar características que sem uma aplicação visual seria muito difícil, como à relação interna dos criptogramas dentro do seu grupo. A Figura 11 ilustra o componente sendo utilizado.

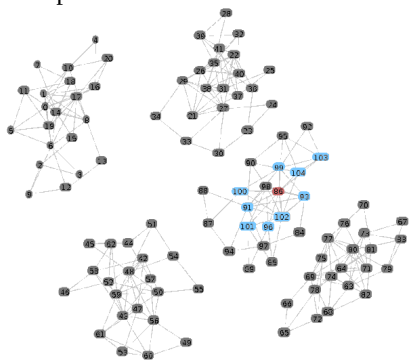


Figura 11. Componente de visualização dos clusters.

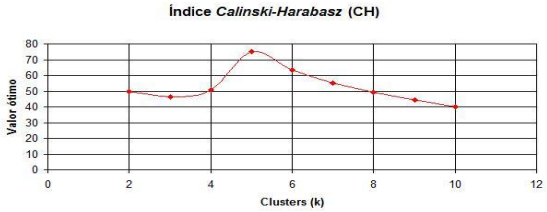


Figura 12. Componente de visualização da função CH.

O segundo componente tem como objetivo ilustrar o comportamento da função de avaliação CH utilizada pelo algoritmo genético, dando maior compreensão da qualidade dos grupos para o agrupamento automático,

pois com o gráfico é fácil identificar o joelho máximo global da função correspondente ao agrupamento adequado. A Figura 12 ilustra o componente sendo utilizado.

4. Experimentos

4.1. Experimentos de (CARVALHO, 2006)

(CARVALHO, 2006) realizou vários experimentos com o objetivo de agrupar criptogramas em função da chave de cifrar. Desta forma, fixava um algoritmo criptográfico e variava o conjunto de chaves. Dentre seus experimentos, o mais interessante foi quando ele introduziu uma maior variabilidade na coleção de teste: 300 textos de dez tamanhos distintos foram cifrados com o DES utilizando 20 chaves pseudo-aleatórias (15 textos para cada chave). Os tamanhos, em bytes, das mensagens foram 64, 128, 256, 384, 512, 1024, 2048 e 3072. O objetivo de variar o tamanho dos textos foi analisar o comportamento dos grupos resultantes do agrupamento, verificando se os criptogramas maiores conseguem corrigir os erros de revocação provocados pela existência de textos menores.

Os resultados comprovam que usando criptogramas de tamanho menores que 512 bytes, nem sempre é possível obter a revocação desejada. Entretanto, em dois casos, os grupos obtidos estavam completos (revocação igual a um), e em outros dez, apenas um único criptograma foi isolado do seu grupo. Assim, os resultados desse experimento podem ser considerados satisfatórios.

Outra contribuição de (CARVALHO, 2006) foi o estabelecimento em bytes da menor amostra a ser utilizada de forma que o agrupamento dos criptogramas de acordo com a chave utilizada fosse realizado com sucesso. Nesse experimento, em cada etapa foram extraídos da Bíblia 30 textos de tamanhos iguais. Estes textos foram cifrados, com o DES, usando 50 chaves pseudo-aleatórias. Assim, foi possível verificar um tamanho recomendável para que o agrupamento seja realizado com sucesso.

Como esperado, os valores de precisão foram sempre iguais a um, indicando que todos os grupos contêm apenas documentos cifrados com uma mesma chave. Já para a revocação, os resultados só foram perfeitos, para criptogramas de pelo menos 512 bytes com amostras de 1500 criptogramas. Nesses tamanhos, conforme confirmado pelos valores de revocação, todos os textos cifrados com uma mesma chave foram colocados no mesmo grupo.

Em tamanhos menores que 512 bytes, textos cifrados com uma mesma chave, em alguns momentos, pertenceram a grupos distintos. Esses resultados comprovam a influência do tamanho da mensagem no sucesso do processo de agrupamento. Desta forma, pode ser concluído que a menor amostra utilizada com sucesso por (CARVALHO, 2006) teve uma configuração de 1500 criptogramas de 512 bytes totalizando uma amostra de 768000 bytes.

4.2. Experimentos de (SOUZA, 2007)

(SOUZA, 2007) realizou vários experimentos, por este motivo uma de suas contribuições foi a construção de um repositório de criptogramas para teste de criptoanálise. Outra contribuição adicional ao trabalho de (Carvalho, 2006) foi que além de agrupar pela chave criptográfica, (SOUZA, 2007) também conseguiu agrupar pelo algoritmo criptográfico utilizado. Neste experimento foi utilizada uma amostra de 30 textos da Bíblia de 4 Kbytes, uma chave e cinco algoritmos criptográficos distintos.

O resultado deste experimento mostrou o sucesso do modelo utilizado por (SOUZA, 2007) na identificação de criptogramas pela chave ou pelo algoritmo criptográfico. Outra contribuição foi a redução do tamanho total da amostra em comparação com a amostra utilizada por (CARVALHO, 2006), visto que foram utilizados 150 criptogramas de 4 Kbytes totalizando uma amostra de 614400 bytes.

Adicionalmente com a implementação de outros experimentos, (SOUZA, 2007) pode testar com a mesma configuração anterior as diversas medidas de similaridade. Foram testadas medidas de similaridade e distância, como: Co-seno, Dice, Jaccard, Overlap, Simple-matching, distância Euclidiana, distância Manhattan, distância Canberra e distância Bray-Curtis; e métodos de agrupamentos hierárquicos aglomerativos, como: single-link, complete-link e group-average link, confirmando o melhor desempenho da medida de similaridade co-seno e do método single-link dentro do contexto proposto.

4.3. Experimento de (TORRES ET AL, 2010)

Os experimentos realizados por (TORRES et al, 2010) tiveram como principal contribuição o agrupamento automatizado dos criptogramas. Outra contribuição importante deste trabalho foi a verificação da influência do tipo texto em claro utilizado para a construção dos criptogramas. Para verificar esta característica, os ensaios utilizaram 30 textos pseudo-aleatórios de 4 Kbytes, uma chave e cinco algoritmos criptográficos, totalizando uma amostra de 150 criptogramas de 4 Kbytes cada.

Com o resultado deste experimento foi observado que o texto em claro não influencia na transmissão de "assinatura" dos algoritmos criptográficos, posto que os grupos formados apresentavam revocação e precisão iguais a 1. O mesmo experimento foi repetido com objetivo de verificar se o comportamento se repetia para as chaves utilizadas, logo foi fixado um algoritmo criptográfico e foram utilizadas cinco chaves diferentes. O resultado deste experimento também foi satisfatório, concluindo que tanto os algoritmos quanto as chaves utilizadas no processo de cifragem transmitem "assinaturas" independentemente do tipo de texto em claro utilizado.

5. Conclusão

Os testes estatísticos propostos pelo NIST têm como objetivo a identificação de padrões nos criptogramas gerados pelos algoritmos ensaiados e, assim, testar se os mesmos são bons geradores de números pseudo-aleatórios. Entretanto, conforme pôde ser visto ao longo deste trabalho, no caso dos algoritmos finalistas do processo para a escolha do AES, esses testes não foram suficientes para impedir a identificação do algoritmo criptográfico a partir de padrões detectados nos criptogramas, com a utilização de técnicas diferentes daquelas propostas pelo NIST. Neste artigo foram apresentadas quatro técnicas: Agrupamento Hierárquico, Redes Neurais, Algoritmos Genéticos e Grafos, as quais foram capazes de separar corretamente os criptogramas gerados pelos algoritmos finalistas do concurso do AES: MARS, RC6, Rijndael, Serpent e Twofish; o que leva a identificação desses algoritmos a partir dos criptogramas gerados pelos mesmos. A identificação relatada demonstra a existência de _assinaturas_ nos criptogramas, as quais são decorrentes das transformações realizadas pelos algoritmos criptográficos ou da mudança de transformação no algoritmo provocada pela chave utilizada na criptografia. As técnicas apresentadas contribuem principalmente na identificação correta de cifras modernas, apresentado resultados satisfatórios. O modo de operação utilizado nos experimentos foi o ECB (Electronic Codebook). A justificativa para a utilização desse modo está no fato de que os algoritmos devem ter forca suficiente para resistir a ataques sob a condição de "pior caso".

Os resultados dos trabalhos apresentados neste artigo sugerem como pesquisas futuras: a) a identificação e separação de classes de chaves para um mesmo algoritmo; b) estudos para modificar as transformações matemáticas nos algoritmos criptográficos testados neste artigo para que, mesmo em modo ECB, estes não propaguem informações dos textos claros para os criptogramas gerados.

6. Referências Bibliograficas

[1] Soto, J. and Bassham, L, Randomness Testing of the Advanced Encryption Standard Candidates Algorithms, NIST Internal Report, 2000.

[2] Murphy, S, The Power of NIST's Statistical Testing of AES Candidates, Information Security Group, Royal Holloway, University of London, 2000.

[3] Knudsen, L.R. and Meier, W, Correlations in RC6 with a Reduced Number of Rounds, Proceedings of the 7th International Workshop on Fast Software Encryption, 2000.

[4] Carvalho, C. A. B, O uso de técnicas de recuperação de informações em criptoanálise, Instituto Militar de Engenharia, 2006.

[5] Souza, W. A. R, Identificação de padrões em criptogramas usando técnicas de classificação de textos, Instituto Militar de Engenharia, 2007.

[6] Souza, W. A. R et al, Método de Agrupamento de Criptogramas em Função das Chaves de Cifrar, IV Workshop em Algoritmos e Aplicações de Mineração de Dados (SBBD/SBES), 2008.

[7] Dileep, A. D and Sekhar, C. C, Identification of block ciphers using support vector machines, International Joint Conference on Neural Networks, 2006.

[8] Ueda, T. K e Terada, R, Uma Versão Mais Forte do Algoritmo RC6 contra a criptoanálise, VII Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais, 2007.

[9] Nagireddy, A Pattern Recognition Approach to Block Cipher Identification, Thesis (Master of Technology) - Indian Institute of Technology Kanpur, 2008.

[10] Torres, H. R and Oliveira, A. G and Xexéo, M. A. J and Souza, R. A. W and Linden, R. Identificação de chaves e algoritmos criptográficos utilizando Algoritmo Genético e Teoria dos Grafos, 9th International Information and Telecommunication Technologies Symposium, 2010.

[11] Souza, R. A. W and Carvalho, V.A.L and Xexéo, M. A. J.Mecanismo Identificador para n-cifras de bloco, 9th International Information and Telecommunication Technologies Symposium, 2010.

[12] Goldschmidt, Ronaldo and Passos, Emmanuel. Data mining: um guia prático. Rio de Janeiro: Elsevier, 2005.

[13] Calinski, R.B., e Harabasz, J., A Dendrite Method for Cluster Analysis. Communications in Statistics, 3(1), 1-27, 1974.

[14] Maheshwari, Pooja. Classification of Ciphers. Thesis (Master of Technology) - Indian Institute of Technology Kanpur, 2001.

[15] Chandra, G. Classification of Modern Ciphers. Thesis (Master of Technology) - Indian Institute of Technology Kanpur, 2002.

[16] Rao, M. B. Classification of RSA and IDEA Ciphers. Thesis (Master of Technology) - Indian Institute of Technology Kanpur, 2003.

[17] Saxena, G. Classification of Ciphers Using Machine Learning. Thesis (Master of Technology) - Indian Institute of Technology Kanpur, 2008.

[18] Laskari, E.C and Melatiou, G.C and Stamatiou, Y.C and Vrahatis, M.N. Cryptography and Cryptanalysis through Computational Intelligence. Computational Intelligence in Information Assurance and Security. Studies in Computational Intelligence, 2007, Volume 57, 1-49.

[19] Albassal, A.M.B andWahdan, A.M. Neural network based cryptanalysis of a feistel type block cipher. In: Proceedings 2004 International Conference on Electrical, Electronic and Computer Engineering, 2004, pp. 231 - 237.

[20] Dileep, A. D and Swapna, S and Sekhar, C. C and Kant, S and Saxena, P.K. Decryption of Feistel Type Block Ciphers using Hetero-Association Model. In: Proceedings XIV National Conference on Communications, Mumbai, India, 2008.

[21] Rao, K.V.S and Krishna, M.R and Baba, D.B. Cryptanalysis of a Feistel Type Block Cipher by Feed Forword Neural Network Using Right Sigmoidal Signals. In: International Journal of Soft Computing, volume 4, issue 3, 2009, 131 - 135.

[22] Yates, R.B. e Neto, B. R. (1999), Modern information retrieval. Addison Wesley.

[23] Fung, B. C. M., Wang, K. e Ester M. (2003), Hierarchical document clustering using frequent itemsets. Proceedings of the SIAM International Conference on Data Mining, San Francisco.