MINISTÉRIO DA DEFESA
EXÉRCITO BRASILEIRO
SECRETARIA DE CIÊNCIA E TECNOLOGIA
INSTITUTO MILITAR DE ENGENHARIA
CURSO DE MESTRADO EM SISTEMAS E COMPUTAÇÃO

GLÁUCIO ALVES DE OLIVEIRA

A APLICAÇÃO DE ALGORITMOS GENÉTICOS NO RECONHECIMENTO DE PADRÕES CRIPTOGRÁFICOS

Rio de Janeiro
2011

INSTITUTO MILITAR DE ENGENHARIA

GLÁUCIO ALVES DE OLIVEIRA

A APLICAÇÃO DE ALGORITMOS GENÉTICOS NO RECONHECIMENTO DE PADRÕES CRIPTOGRÁFICOS

Dissertação de Mestrado apresentada ao Curso de Mestrado em Sistemas e Computação do Instituto Militar de Engenharia, como requisito parcial para obtenção do título de Mestre em Sistemas e Computação.

Orientador: Prof. José Antonio Moreira Xexéo - D.Sc

Rio de Janeiro
2011

c2011

INSTITUTO MILITAR DE ENGENHARIA
Praça General Tibúrcio, 80-Praia Vermelha
Rio de Janeiro-RJ CEP 22290-270

Este exemplar é de propriedade do Instituto Militar de Engenharia, que poderá incluí-lo em base de dados, armazenar em computador, microfilmar ou adotar qualquer forma de arquivamento.

É permitida a menção, reprodução parcial ou integral e a transmissão entre bibliotecas deste trabalho, sem modificação de seu texto, em qualquer meio que esteja ou venha a ser fixado, para pesquisa acadêmica, comentários e citações, desde que sem finalidade comercial e que seja feita a referência bibliográfica completa.

Os conceitos expressos neste trabalho são de responsabilidade do autor e do orientador.

Oliveira, G. A.
A aplicação de Algoritmos Genéticos no Reconhecimento de Padrões Criptográficos/ Gláucio Alves de Oliveira.
– Rio de Janeiro: Instituto Militar de Engenharia, 2011.
xxx p.: il., tab.

Dissertação (mestrado) – Instituto Militar de Engenharia – Rio de Janeiro, 2011.

1.Reconhecimento de Padrões. 2. Classificação. I. Título. II. Instituto Militar de Engenharia.

CDD 629.892

INSTITUTO MILITAR DE ENGENHARIA

GLÁUCIO ALVES DE OLIVEIRA

A APLICAÇÃO DE ALGORITMOS GENÉTICOS NO RECONHECIMENTO DE PADRÕES CRIPTOGRÁFICOS

Dissertação de Mestrado apresentada ao Curso de Mestrado em Sistemas e Computação do Instituto Militar de Engenharia, como requisito parcial para obtenção do título de Mestre em Sistemas e Computação.

Orientador: Prof. José Antonio Moreira Xexéo - D.Sc

Aprovada em 17 de Janeiro de 2011 pela seguinte Banca Examinadora:

Prof. José Antonio Moreira Xexéo - D.Sc do IME - Presidente

Cel(R/1). Paulo Afonso Lopes da Silva - Ph.D, do IME

Prof. Luis Alfredo Vidal de Carvalho - DSc, da UFRJ/COPPE

Prof. Ricardo Linden - DSc, da CEPEL/ELETROBRÁS

Rio de Janeiro
2011

"Aos meus pais, minha fortaleza."

AGRADECIMENTOS

A Deus por fortalecer-me nos momentos difíceis de minha singradura.

A minha família, em especial aos meus pais, Oscar Alves de Oliveira e Vanir Figueiredo Alves de Oliveira, pelo amor e apoio em toda minha vida.

A Renata da Silva Ferrarezi pelo seu amor incondicional, compreensão e paciência nos muitos momentos em que estive ausente.

Ao meu Professor José Antônio Moreira Xexéo, pela orientação precisa e confiança depositada em minha pessoa no desenvolvimento deste trabalho.

Ao Capitão-de-Corveta William Augusto Rodriguês de Souza, pela co-orientação e atenções dadas no Centro de Análises de Sistemas Navais ao longo de todo o curso ministrado no IME.

Ao Professor Paulo Afonso Lopes da Silva, que muito contribuiu para o enriquecimento do meu aprendizado.

Aos meus colegas de mestrado pelo convívio harmonioso, especialmente a Renato Hidaka Torres, pela cordialidade e companheirismo nos momentos difíceis na transposição dos vários obstáculos no transcorrer do curso.

A Marinha do Brasil e ao Exército Brasileiro pela oportunidade.

A todos os professores e funcionários do Departamento de Engenharia de Sistemas (SE/8) do Instituto Militar de Engenharia pela convivência harmoniosa e ambiente adequado aos trabalhos de pesquisa.

A todas as pessoas que contribuíram direta ou indiretamente com o desenvolvimento desta dissertação de mestrado, tenha sido por meio de críticas, idéias, apoio, incentivo ou qualquer outra forma de auxílio.

Gláucio A. de Oliveira

"Vês um homem diligente em seu trabalho? Ele será posto a serviço de reis..."

Provérbios, 22:29.

SUMÁRIO

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SÍMBOLOS

ABREVIATURAS

AES	-	*Advanced Encryption Standard*
CBC	-	*Cipher Block Chaining*
ECB	-	*Eletronic Codebook*
IME	-	*Instituto Militar de Engenharia*
NIST	-	*National Institute of Standards and Tecnology*

RESUMO

O princípio fundamental das cifras de bloco é a geração de criptogramas com uma distribuição que não estabeleça correlação com os dados de entrada (textos claros ou chaves). Estudos em modernos processos de criptografia evidenciam indícios de padrões de "assinatura" associados ao tipo de algoritmo ou a chave utilizados no processo de cifragem.

Em um ataque somente com texto cifrado (*Ciphertext-only attack*), são poucas as informações disponíveis para um criptoanalista. Existe a necessidade de conhecer pelo menos o algoritmo criptográfico utilizado na cifragem. Neste contexto, este trabalho descreve o uso do Algoritmo Genético (AG) como um modelo de ferramenta para o agrupamento de criptogramas gerados por algoritmos criptográficos certificados pelo **NIST** (*National Institute Standard Tecnology*). Os ensaios realizados com o Algoritmo Genético realizaram o agrupamento de criptogramas gerados por algoritmos cifrantes distintos e pelo mesmo algoritmo cifrante com chaves distintas. Adicionalmente, uma técnica de classificação conhecida como *Template Matching* foi utilizada com o Algoritmo Genético. O Algoritmo Genético agrupador que utiliza a técnica de classificação é denominado neste trabalho como "Algoritmo Genético modelado".

ABSTRACT

The basic principle of block ciphers is the generation of cryptograms with a distribution that does not establish a correlation with the input data (plaintext or keys). Studies in modern methods of encryption show evidence of "signature" associated with the type of algorithm or the key used in the encryption process.

In a ciphertext-only attack, there is little information available to a cryptanalyst, who needs to know at least the cryptographic algorithm used in encryption. In this context, this paper describes the use of a Genetic Algorithm (GA) as a tool for clustering cryptograms generated by cryptographic algorithms certified by NIST (National Institute Standard Technology). Tests with the Genetic Algorithm performed by the clustered cryptograms both generated by different algorithms and encrypting the same encrypted algorithm with different keys. Additionally, a technique known as "Template Matching" was used with Genetic Algorithm for classification .

The GA that uses the technique of classification is called in the work "Modeled Genetic Algorithm".

1 INTRODUÇÃO

A criptografia tem como objetivo a conversão de textos em claro em criptogramas. Esses criptogramas são gerados por cifras de blocos ou de fluxo cujo objetivo é a eliminação da correlação entre os dados de entrada (chave/cifra/texto claro) com os de saída (criptograma). Todas as informações são convertidas em sequências de números binários integradas a uma matemática computacional cujo principio fundamental é uma forte aleatorização com o propósito de gerar criptogramas com uma distribuição uniforme de caracteres.

Vários estudos e ensaios em algoritmos de criptografia evidenciaram indícios de uma "assinatura" associada ao tipo de cifra ou da chave em uso. Na literatura, por exemplo, observam-se pesquisas e trabalhos que têm desenvolvido métodos e técnicas baseadas na Inteligência Artificial, Técnicas de Agrupamento e Lógica Fuzzy em cifras de bloco tais como o AES e DES, todas com o objetivo de agrupar criptogramas, consequência da correlação dos textos cifrados com os algoritmos (cifras) ou chaves que os geraram.

1.1 MOTIVAÇÃO

(CARVALHO, 2006) e (SOUZA, 2007) detectaram indícios de padrões em criptogramas gerados pelo DES e AES, correlacionando-os com o tipo de chave ou algoritmo que os geraram, independentemente do conhecimento das operações dos processos internos das cifras. A detecção de tais padrões criptográficos realizou-se com técnicas de "Agrupamento Hierárquico" que obteve resultados satisfatórios no agrupamento de criptogramas pelo tipo de chave usada. Entretanto, ressalta-se que para a realização correta das operações de agrupamento, era necessário inicialmente ter o conhecimento do número exato de grupos distintos que poderiam ser formados. Adicionalmente, por meio de Redes Neuronais (Mapa de Kohonen), também foi feita uma tentativa de classificação de criptogramas gerados por diferentes tipos de cifra ou chave em uso. O mapa gerado agrupou no mesmo grupo criptogramas gerados por algoritmos criptográficos distintos. Portanto, os resultados dessa classificação mostraram-se insatisfatórios.

Um bloco com o comprimento de n bits tem 2^n possibilidades de entrada em um sistema criptográfico proporcionando 2^n possibilidades de saída. O elevado número de

possibilidades de entrada dificulta a análise isolada do criptograma para fins de uma eficiente e possível decriptografia não-autorizada.

O Algoritmo Genético é uma técnica computacional usada em problemas de busca de soluções em espaços intratavelmente grandes. O seu uso justifica-se na tentativa de detectar padrões em criptogramas e agrupá-los em função dos algoritmos ou das chaves que os geraram. Associado ao Algoritmo Genético, existe um método de classificação chamado *Template Matching* cujo objetivo é classificar isoladamente criptogramas, "a priori" desconhecidos. A utilização de um Algoritmo Genético associado um método de classificação visa a incrementar e melhorar a busca de soluções. Destarte, esta dissertação foi motivada pelas seguintes metas:

a) Melhorar o desempenho do agrupamento dos criptogramas realizado por (CARVALHO, 2006) e (SOUZA, 2007);

b) Agrupar e classificar os criptogramas gerados pelos algoritmos criptográficos finalistas do AES, sem a necessidade de conhecer a quantidade de algoritmos distintos em análise;

c) Propor métodos de certificação de segurança criptográfica nas Forças Armadas brasileiras como o objetivo de testar e avaliar a qualidade de algoritmos criptográficos criados no âmbito interno militar ou adquiridos comercialmente;

d) Contribuir para o aperfeiçoamento dos critérios e testes de certificação de segurança estabelecido pelo NIST (*National Institute of Standards and Tecnology*);

e) Contribuir para a fase preliminar do ataque por só-texto-ilegível (*Ciphertext-only attack*).

A tarefa de agrupamento será realizada pelo Algoritmo Genético e a classificação pela utilização da técnica de Reconhecimento de Padrões conhecida como *Template Matching*. Nesta dissertação será usado o termo "Algoritmo Genético modelado" que corresponderá ao Algoritmo Genético agrupador, associado à técnica utilizada do *Template Matching*. A figura 1.1 ilustra, sucintamente, o funcionamento do sistema de agrupamento e classificação utilizado neste trabalho.

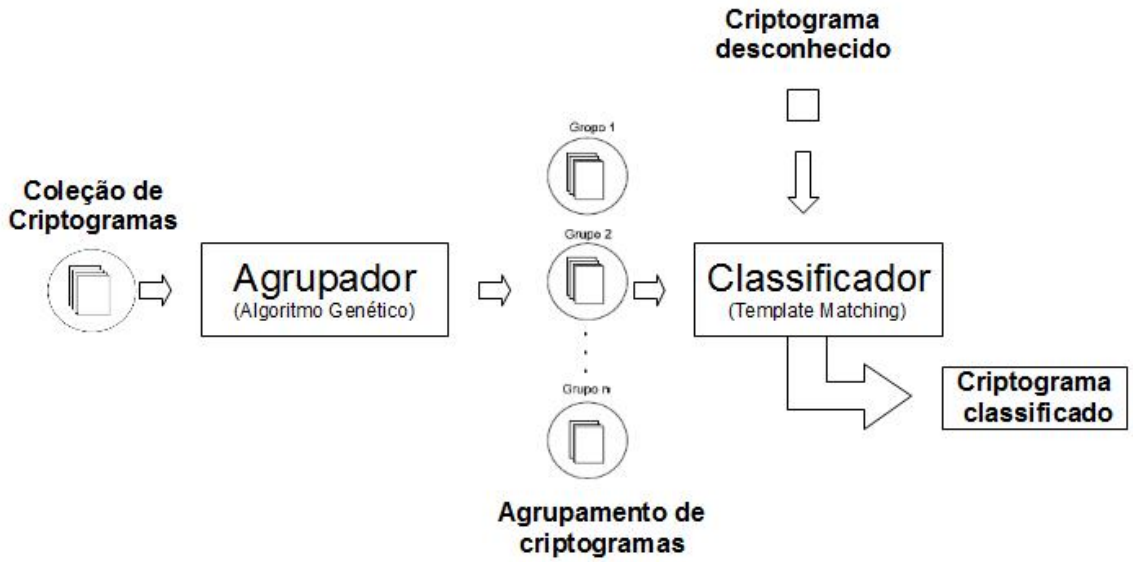


FIG. 1.1: Sistema agrupador e classificador. Identificação do tipo de algoritmo criptográfico ou chave em uso de um criptograma desconhecido.

1.2 CARACTERIZAÇÃO DO PROBLEMA

Os algoritmos criptográficos utilizados comercialmente como o DES, o RSA e, recentemente, o AES passaram por intensivas avaliações tanto do meio acadêmico quanto de órgãos como o NIST. O objetivo dessas avaliações, tanto de algoritmos quanto dos produtos criptográficos, é garantir um mínimo de segurança no uso desses artefatos.

O processo de certificação utilizado pelo NIST, quando da concorrência que elegeu o AES como padrão, baseou-se em testes estatísticos. Os relatórios desses ensaios induzem à conclusão de que a métrica utilizada para qualificar o nível de segurança dos algoritmos testados é função dos níveis de aleatoriedade determinados nos diversos ensaios.

(SOTO, 2000) ao relatar os ensaios realizados com os finalistas do AES (Mars, RC6, Rijndael, Serpent e Twofish), estabeleceu duas assertivas: a) "embora se acredite que os cinco algoritmos finalistas geram sequências aleatórias, os testes foram realizados para mostrar que há evidências empíricas para apoiar essa convicção"; b) "os resultados sugerem que, apesar de anomalia detectada no Serpent, é lícito afirmar que todos os algoritmos parecem não ter desvios da aleatoriedade detectáveis". Quase simultaneamente, (MURPHY, 2000) discutiu a metodologia utilizada nesses ensaios e, embora não questionasse a certificação, concluiu seu relatório com diversas restrições, entre elas: a) "as hipóteses de

teste não são suficientemente claras, gerando interpretações diferentes para os resultados"; b) "testes estatísticos equivalentes, realizados com os mesmos dados, não geram, necessariamente, os mesmos resultados". Havia, portanto, controvérsias se não pela qualidade dos algoritmos, pelo menos pela completude e propriedade dos ensaios realizados. Mais recentemente, diversos pesquisadores levantaram outras questões em relação aos resultados apresentados por (SOTO, 2000).

Diversos tipos de testes detectaram padrões nos criptogramas gerados pelos algoritmos submetidos aos testes do NIST. Estes testes permitiram distinguir criptogramas tanto por algoritmo quanto por chave de origem. Esses resultados são importantes porque explicitam indícios da transmissão de "assinaturas" dos algoritmos e das chaves aos criptogramas. (KNUDSEN, 2000) realizou o chamado "ataque de distinção", para criptogramas gerados pelo algoritmo RC6 com o auxílio da estatística do *qui-quadrado*(χ^2). (CARVALHO, 2006), (SOUZA, 2007) agruparam criptogramas gerados pelas cifras de blocos DES, AES e RSA, em função da chave, com diversos tamanhos de criptogramas e chaves. (DILEEP, 2006) usou máquinas de vetor de suporte (SVM) para identificar as cifras de bloco DES, Triple DES, Blowfish, AES e RC5, a partir de conjuntos de criptogramas gerado por ele. (UEDA, 2007) propôs um método para fortalecer o algoritmo RC6, como uma contramedida ao ataque com a estatística do *qui-quadrado*. (NAGIREDDY, 2008) desenvolveu métodos de histograma e de predição de bloco, técnicas de expansão de dados e técnicas baseadas em ataques secundários para identificação das cifras de bloco DES, AES, Blowfish, Triple DES e RC5. Os resultados desses trabalhos reforçaram a hipótese da existência de "assinaturas" transmitidas para os criptogramas pelas chaves e pelos próprios algoritmos.

Em suma, um dos requisitos que pode ser utilizado para a garantia da confiabilidade de um algoritmo criptográfico seria a inexistência de "assinaturas" nos criptogramas gerados por estes algoritmos. Portanto, o reconhecimento de padrões detectados nos criptogramas gerados por um mesmo tipo de cifra ou chave acarreta em um forte questionamento sobre o nível de aleatoriedade da saída de uma cifra de bloco.

1.3 ORGANIZAÇÃO DA DISSERTAÇÃO

No capítulo 2 é apresentada uma sucinta revisão bibliográfica comentada sobre os conceitos básicos da criptografia e do Reconhecimento de Padrões aplicados neste trabalho.

O capítulo 3 relata o desenvolvimento da fase de pré-processamento para o tratamento dos criptogramas que necessitam serem modelados em uma estrutura de dados de entrada para o Algoritmo Genético.

O capítulo 4 descreve a modelagem do Algoritmo Genético para a tarefa de agrupamento de criptogramas.

No capítulo 5 é descrita a metodologia de classificação associada ao Algoritmo Genético com o objetivo de classificar criptogramas, a priori desconhecidos.

No capítulo 6 são descritos os experimentos e feitas análises e avaliações dos resultados apresentados.

O capítulo 7 compara o desempenho do Algoritmo Genético modelado com outras técnicas modeladas utilizadas para o agrupamento e classificação de criptogramas.

No capítulo 8 há a conclusão, as contribuições desta dissertação e a perspectiva de trabalhos.

2 REVISÃO BIBLIOGRÁFICA COMENTADA

2.1 INTRODUÇÃO

Este capítulo contém uma síntese de criptografia e Reconhecimento de Padrões (RP). Enquanto a criptografia está direcionada à segurança da informação, o Reconhecimento de Padrões, no contexto da criptologia [1], está aplicado na verificação da eficiência dos algoritmos criptográficos [2] no que tange a produção de textos ilegíveis com um alto nível de distribuição aleatória de *caracteres*. (DILEEP, 2006) considera o RP em textos ilegíveis gerados por cifras como um tipo de criptoanálise [3].

2.1.1 CRIPTOGRAFIA

2.1.1.1 DEFINIÇÃO

A criptografia é a ciência que trata do projeto, análise e implantação de algoritmos criptográficos com o propósito de garantir o segredo das mensagens. Um algoritmo criptográfico é uma função que utiliza técnicas matemáticas para transformar um texto originalmente legível (texto claro) em um criptograma [4]. O processo de conversão de um texto claro em um criptograma é chamado de cifragem. O processo inverso para converter um criptograma em um texto claro é chamado de decifragem. Os processos de cifrar e decifrar englobam, cada um, a execução de duas funções descritas por algoritmos que são combinados com uma chave [5]. A figura 2.1 exemplifica um modelo genérico de comunicações utilizando as funções de cifrar e decifrar.

[1] É o estudo das comunicações seguras, que abrange tanto a criptografia quanto a criptoanálise.

[2] Os termos algoritmo criptográfico e cifra são intercambíaveis nesta dissertação.

[3] É o ramo da criptologia que trata da "quebra" de uma cifra para recuperar informações (STALLINGS, 2008)

[4] Texto ilegível

[5] É o elemento que transforma o algoritmo de cifragem geral num método específico de cifragem. O inimigo pode saber qual é o algoritmo de cifragem sendo usado pelo remetente e o destinatário da mensagem, mas ele não pode conhecer a chave.

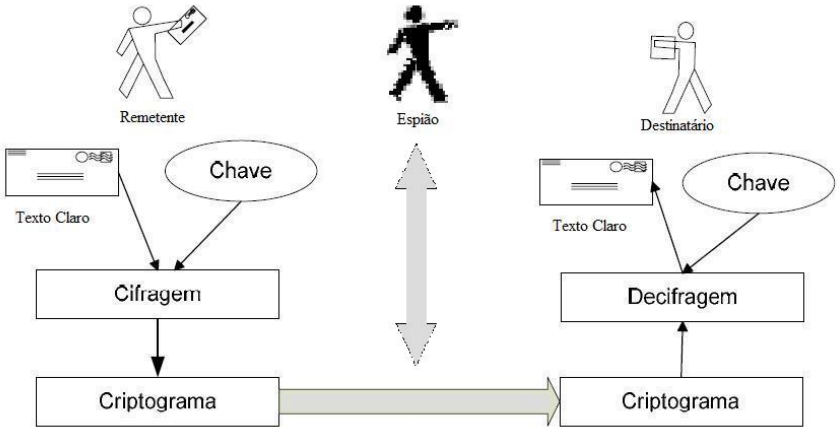


FIG. 2.1: Modelo genérico de comunicações (SOUZA, 2007).

Assim, um sistema criptográfico clássico é uma quíntupla (P, C, K, E, D), no qual as seguintes condições são satisfeitas (STINSON, 2006):

1. P é um conjunto finito de possiveis textos claros;
2. C é um conjunto finito de possíveis textos cifrados;
3. K é um conjunto finito de possíveis chaves;
4. Para cada $k \in K$, existe uma função de cifrar $e_k \in E$ e uma correspondente função de decifrar $d_k \in D$. Cada $e_k : P \to C$ e $d_k : C \to P$ são funções tal que $d_k(e_k(x)) = x$ para todo texto claro $x \in P$.

O remetente ou origem produz uma mensagem em texto claro formada pela concatenação de blocos binários, $x = x_1 x_2 \cdots x_n$. Para qualquer inteiro $n \geqslant 1$, onde para cada bloco tem-se $x_i \in P$, $1 \leqslant i \leqslant n$. Cada bloco x_i é criptografado usando a função de cifrar e_k que utiliza uma chave predeterminada k. Então, o remetente calcula $y = e_k(x_i)$, $1 \leqslant i \leqslant n$ resultando no texto cifrado $y = y_1 y_2 \cdots y_n$.

Quando o destinatário recebe a mensagem cifrada y, a função de decifrar d_k é usada para obtenção da mensagem de texto claro x. A função de cifrar deve ser injetiva. Para todas as mensagens x_1, x_2 ϵ P, que satisfaçam a condição $y = e_k(x_1) = e_k(x_2)$, tem-se $x_1 = x_2$.

2.1.1.2 A IMPORTÂNCIA DA CRIPTOGRAFIA

Nos dias atuais, chamadas telefônicas deslocam-se entre satélites e mensagens eletrônicas trafegam por vários computadores. Ambas os modos de comunicação podem ser interceptadas, ameaçan-do a privacidade. De modo semelhante, à medida que mais negócios são realizados por meio da rede mundial de computadores, devem ser instalados mecanismos de proteção para a segurança das pessoas físicas e jurídicas. A criptografia é um meio de proteger a privacidade e garantir o sucesso do mercado digital. Os defensores das liberdades civis começam a pressionar pelo uso generalizado da criptografia de modo a proteger a privacidade dos indivíduos. Ao lado deles fica a comunidade dos negócios, que precisa de uma criptografia para proteger suas transações no mundo em rápido crescimento do comércio via Internet (SINGH, 2008). Desse modo, a segurança da informação é um problema importante.

(STALLINGS, 2008) identifica os requisitos necessários para se estabelecer uma comunicação segura:

- Confidencialidade - Garantia de que as informações armazenadas em um sistema de computacional sejam acessadas somente por indivíduos autorizados;
- Autenticidade - Garantia de que os indivíduos participantes de uma comunicação sejam corretamente identificados.
- Integridade - Garantia de que a informação processada ou transmitida chegue ao seu destinatário exatamente da mesma forma em que partiu do remetente. A informação não contém modificação, inserção, exclusão ou repetição no trâmite de comunicação entre envio e recepção; e
- Irretratabilidade - Garantia de que nem o remetente e nem o destinatário das informações possam negar, posteriormente, sua transmissão, recepção ou posse.

2.1.1.3 TIPOS DE CRIPTOGRAFIA

Existem vários algoritmos criptográficos que podem ser classificados, quanto ao segredo da chave para cifrar, em simétricos ou assimétricos. Assim, existem dois tipo de criptografia:

- Criptografia simétrica ou criptografia de chave secreta;
- Criptografia assimétrica ou criptografia de chave pública.

CRIPTOGRAFIA SIMÉTRICA

A criptografia simétrica utiliza a mesma chave para cifrar e decifrar. Esta chave, a rigor, não precisa ser única, pois a chave para decifrar pode ser obtida a partir da chave utilizada para cifrar (DENNING, 1982)(DIFFIE, 1976). A chave deve ser mantida em sigilo e compartilhada entre um remetente e destinatário em um canal seguro de comunicação. A figura 2.2 ilustra um modelo simplificado de criptografia simétrica.

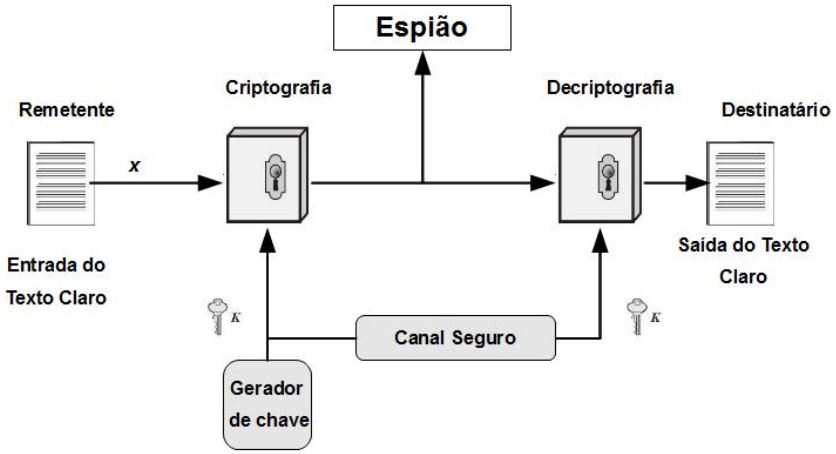


FIG. 2.2: Modelo de criptografia simétrica.

A criptografia simétrica é subdividida em dois tipos de cifras:

- Cifras de fluxo - São cifras que criptografam um fluxo de dados digital um bit ou byte de cada vez.
- Cifras de bloco - São cifras que dividem um texto claro em um blocos de bits e usam uma função especial para misturar um bloco do texto claro com a chave secreta e assim produzir o texto cifrado.

CRIPTOGRAFIA ASSIMÉTRICA

Na criptografia assimétrica existem dois tipos de chaves. Neste par de chaves, uma é usada para a criptografia e a outra, diferente, contudo relacionada, para a decriptografia. É importante ressaltar que qualquer uma das duas chaves relacionadas pode ser usada para a criptografia, com a outra usada para a decriptografia. Estas duas chaves são denominadas chave pública e chave privada. A chave pública é de domínio público e doutrinariamente utilizada para cifrar. Ela é distribuída em uma área ostensiva conhecida como arquivo ou diretório público. A chave privada correspondente, que é utilizada para decifrar, deve permanecer em segredo. As chaves públicas e privadas estão matematicamente correlacionadas entretanto, é computacionalmente inviável [6] para qualquer oponente recuperar a chave privada por meio de acesso à chave pública. Quando um remetente deseja enviar uma mensagem criptografada para um destinatário, ele cifra um texto claro com a chave pública do destinatário. Este ao receber a mensagem cifrada, utiliza a chave privada correspondente para decriptografá-la.

O remetente também poderia enviar uma mensagem criptografada utilizando a sua própria chave privada. Neste caso, o destinatário ao receber a mensagem cifrada, utilizaria a chave pública do remetente para decriptografá-la. Desta forma, como a mensagem enviada foi cifrada usando a chave privada do remetente; então, somente o remetente poderia ter criptografado a mensagem. Portanto, a mensagem criptografada por uma chave privada serve como uma assinatura digital. A figura 2.3 mostra um modelo de criptografia assimétrica.

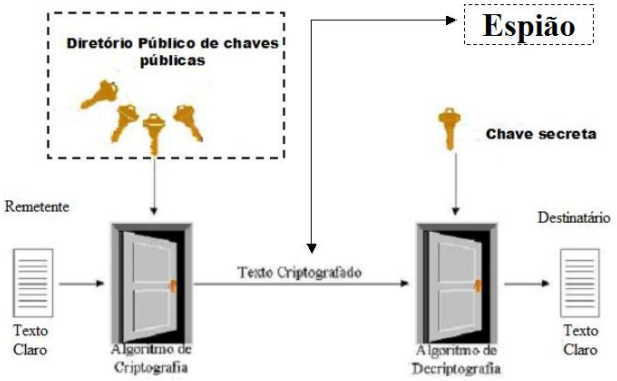


FIG. 2.3: Modelo de criptografia assimétrica.

[6] O tempo e o custo para violar a segurança são altos demais, considerada a capacidade atual dos computadores modernos.

2.1.1.4 DISCUSSÃO

Uma das desvantagens da criptografia simétrica é o gerenciamento na distribuição de chaves. O número de chaves aumenta em uma ordem de grandeza proporcional ao quadrado do número de participantes. Se as comunicações forem dois a dois e se for usada uma chave para cada ligação, então serão necessárias um número k de chaves calculado pela equação 2.1.

$$k = \frac{n(n-1)}{2} \tag{2.1}$$

onde n é o número de participantes.

Deste modo, quando temos um número muito grande de participantes na comunicação, teremos um número grande de chaves a serem gerenciadas no momento de sua transmissão ou no armazenamento, o que trará uma dificuldade adicional para a criptografia simétrica.

Outra desvantagem da criptografia simétrica é a inexistência do requisito de segurança de irretratabilidade. Em virtude do uso de uma única chave tanto para o rementente quanto para o destinatário em um canal de comunicação, não há a confirmação de qual dos dois criptografou a mensagem. Somente é possível garantir a irretratabilidade com a criptografia assimétrica.

Uma desvantagem dos algoritmos de criptografia assimétrica é a velocidade baixa de processamento em virtude do envolvimento de cálculos mais complexos, como exponenciação de valores muito grandes ou cálculos com curvas elípticas (BENITS, 2003) (NAGIREDDY, 2008).

Segundo (RANDALL, 2002) e (STALLINGS, 2008), as cifras de blocos são as mais usadas, principalmente nas aplicações de criptografia simétrica fundamentadas em rede. Destarte, esta dissertação está focalizada nas cifras de bloco finalistas do concurso do AES e suas análises.

2.1.1.5 MODOS DE OPERAÇÃO DE CIFRA DE BLOCO

O modo de operação é uma técnica para aperfeiçoar o efeito de um algoritmo criptográfico sobre uma sequência de blocos de dados ou fluxo de dados. Esses modos de operação

cifram os dados de maneira isolada ou encadeada. Dois deles são sucintamente descritos abaixo. Outros modos de operação pode ser encontrados no NIST [7]. Há também modos de operação que estão sendo desenvolvidos e submetidos [8] à aprovação do próprio NIST.

MODO ELETRONIC CODEBOOK (ECB)

O ECB é o modo mais simples de cifragem. O texto claro é tratado um bloco de cada vez. Cada bloco do texto claro é criptografado usando a mesma chave. Uma característica e desvantagem do modo ECB é que a criptografia de blocos de texto claro idênticos (cifrados com uma mesma chave) produz blocos cifrados idênticos. A vantagem do modo ECB é o paralelismo nas operações de criptografia e decriptografia nos blocos de texto claro e cifrado, respectivamente. Desta forma, há um alto desempenho na velocidade no processo de cifragem e decifragem. As figuras 2.4 e 2.5 ilustram o modo de operação ECB.

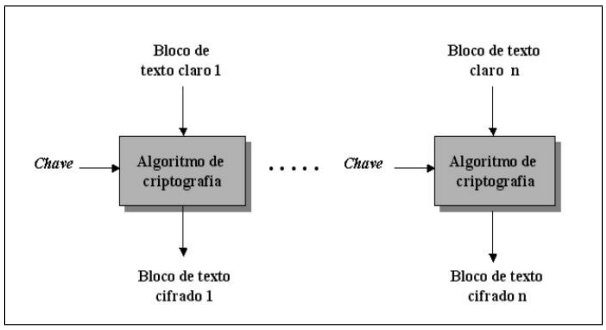


FIG. 2.4: Modo ECB de criptografia .

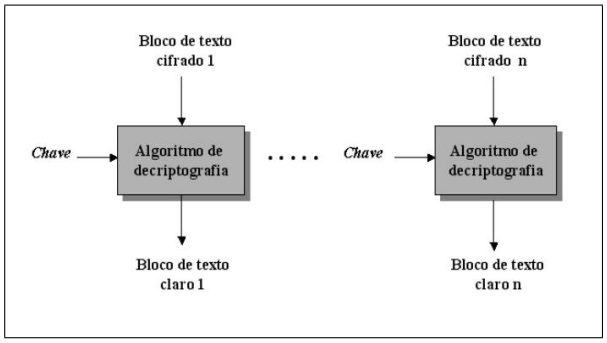


FIG. 2.5: Modo ECB de decriptografia.

[7] http://csrc.nist.gov/groups/ST/toolkit/BCM/current_modes.html
[8] http://csrc.nist.gov/groups/ST/toolkit/BCM/modes_development.html

MODO CIPHER BLOCK CHAINING (CBC)

No modo de criptografia CBC, cada bloco de texto claro é somado, por meio de uma operação ou-exclusivo, com o bloco precedente criptografado. A mesma chave é usada para cada bloco. Um Vetor de Inicialização (VI)[9] é necessário para cifrar o primeiro bloco. Deste modo, há a vantagem de não ocorrer a repetição de blocos cifrados como acontece no modo ECB. A desvantagem no uso do modo CBC é a não existência do paralelismo durante as operações de cifragem e decifragem. Isto pode não ser adequado em operações de criptografia o qual o tempo de processamento criptográfico é um fator crítico.

(NAGIREDDY, 2008) afirma que o Vetor de Inicialização (VI) não necessita ser secreto pelo fato de que o conhecimento dele por parte de algum oponente não compromete um sistema criptográfico. Em contrapartida, (STALLINGS, 2008) diz que o Vetor de Inicialização (VI) deve ser protegido e de conhecimento apenas do remetente e destinatário pois, se um oponente tiver a capacidade de enganar o destinatário da mensagem, convencendo-o a usar um valor diferente para o VI, então o oponente será capaz de inverter os bits selecionados no primeiro bloco de texto claro. As figuras 2.6 e 2.7 ilustram o modo de operação CBC.

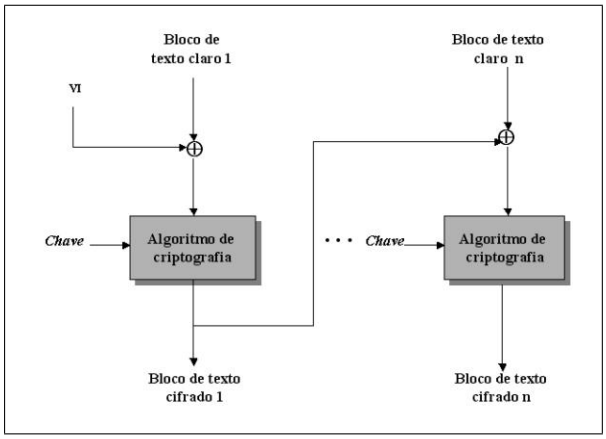


FIG. 2.6: Modo CBC de criptografia.

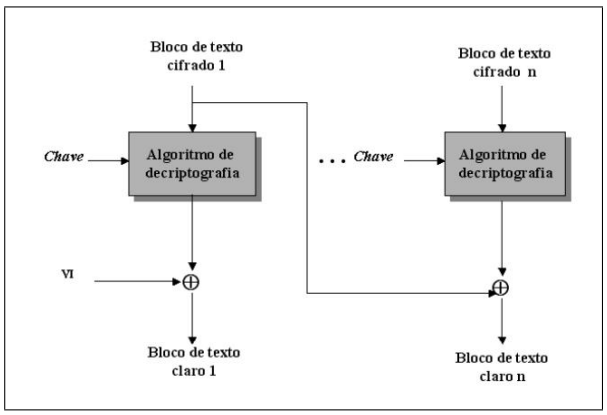


FIG. 2.7: Modo CBC de decriptografia.

[9]Bloco aleatório de dados usado para iniciar a cifragem de blocos de texto claro, quando é utilizada a técnica de criptografia por encadeamento de bloco (CBC).

2.1.1.6 TIPOS DE ATAQUES CRIPTOANALÍTICOS

Esta seção descreve suscintamente os diversos tipos de ataques na criptoanálise fundamentados na quantidade de informação conhecida pelo criptoanalista.

ATAQUE POR SÓ-TEXTO-ILEGÍVEL

Neste tipo de ataque o criptoanalista detém apenas o(s) texto(s) criptografado(s) em mãos. Não há nenhuma outra informação disponível. É considerado como o ataque mais difícil a ser praticado. Nessas circunstâncias, pode ser utilizado o método de força bruta [10] que tenta todas as chaves possíveis em todos os tipos de cifras. Se o espaço de chaves for muito grande, o ataque torna-se computacionalmente inviável. Qualquer sistema de criptografia vulnerável a este tipo de ataque é considerado completamente inseguro (MENEZES, 1996) (BENITS, 2003) (STALLINGS, 2008) (NAGIREDDY, 2008).

Duas importantes tarefas na criptoanálise quando somente o texto cifrado está disponível são: A identificação do método de cifragem e a identificação da chave usada (DILEEP, 2006). Neste contexto, esta dissertação utiliza o Reconhecimento de Padrões, por meio do Algoritmo Genético modelado, na classificação do tipo de cifra que está sendo utilizado e assim, tentar reduzir o esforço empregado pela criptoanálista.

ATAQUE POR TEXTO CLARO CONHECIDO

No ataque por texto claro-conhecido o criptoanalista tem acesso a um texto cifrado e seu correspondente texto claro. Objetivo é recuperar a chave em uso. Este tipo de ataque é pouco provável de acontecer em virtude da necessidade dos pares de textos correlacionados (MENEZES, 1996)(NAGIREDDY, 2008).

ATAQUE POR TEXTO CLARO ESCOLHIDO

Neste tipo de ataque o criptoanalista escolhe o texto claro a ser cifrado, e logo após a sua criptografia, é dado o texto cifrado correspondente que será utilizado para análise. Na escolha de textos claros para serem criptografados, o criptoanalista poderá escolher deliberadamente padrões que poderão revelar a estrutura da chave em uso (MENEZES, 1996)(STALLINGS, 2008).

[10] O ataque por força bruta é uma tentativa de cada chave possível até que seja obtida uma tradução inteligível de texto cifrado para texto claro (STALLINGS, 2008).

ATAQUE POR TEXTO CIFRADO ESCOLHIDO

É similiar ao ataque de texto claro escolhido. O criptoanalista escolhe um texto cifrado juntamente com o seu texto claro decriptografado correspondente, gerado com a chave secreta (NAGIREDDY, 2008)(STALLINGS, 2008).

2.1.2 RECONHECIMENTO DE PADRÕES

2.1.2.1 DEFINIÇÃO

O Reconhecimento de Padrões [11] é uma disciplina científica cujo o objetivo é a classificação de objetos em um número de categorias ou classes [12] (THEODORIDIS, 2009). A tarefa de classificar exige a comparação de um objeto com outros objetos que supostamente pertençam a classes anteriormente definidas. A comparação entre objetos é realizada por meio de uma métrica ou medida de diferença entre eles (CARVALHO, 2005).

No Reconhecimento de Padrões existem dois tipos de classificação, em função da *aprendizagem* ou *treinamento*(*training*) de um sistema classificador:

1 Classificação Supervisionada; e

2 Classificação Não-Supervisionada.

Na classificação supervisionada admite-se o conhecimento a *priori* de classes geradas por padrões detectados em um conjunto de objetos disponíveis denominado "conjunto de treinamento"(*set of training*). Assim, a informação inicial de cada classe rotulada é, posteriormente, comparada com objetos desconhecidos para que os mesmos sejam classificados (rotulados) em alguma classe conhecida.

Na classificação não-supervisionada não há o conhecimento a *priori* de nenhum tipo de classe. Desta maneira, é feita a classificação de um conjunto de objetos em classes desconhecidas a *priori* em número ou forma. O objetivo é descobrir em quantas classes os objetos desconhecidos se distribuem e como são estas classes (CARVALHO, 2005). Por esta razão, por exemplo, a tarefa de agrupamento (*clustering*) é considerada na literatura de Reconhecimentos de Padrões como uma classificação não-supervisionada.

[11] Padrões são as propriedades que possibilitam o agrupamento de objetos semelhantes dentro de uma determinada classe ou categoria, mediante a interpretação de dados de entrada, que permitam a extração das características relevantes desses objetos (TOU, 1981).

[12] Classe é um conjunto de atributos comuns aos objetos de estudo (PRADO, 2008)

Assim, na tarefa de agrupamento não existe nenhum conhecimento a *priori* do conjunto ou coleção de objetos que estão sendo analisados, ao contrário do que ocorre com a classificação supervisionada. Nesta dissertação, são utilizados os termos "classificação" e "agrupamento" para definir a tarefa de "classificação supervisionada" e "classificação não-supervisionada", respectivamente.

A figura 2.8 ilustra a estrutura clássica de um sistema de Reconhecimento de Padrões. Esta estrutura é constituída por dois blocos: um bloco de extração de características, e um classificador.

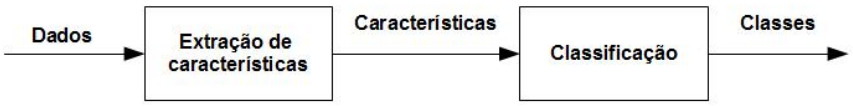


FIG. 2.8: Estrutura típica de um sistema de Reconhecimento de Padrões (MARQUES, 2005).

Segundo (BANDYOPADHYAY, 2007) e (THEODORIDIS, 2009), a fase de extração de características é a tarefa mais importante em um sistema automático de Reconhecimento de Padrões. É nesta fase que os padrões ou características notáveis de um conjunto de objetos são detectados e armazenados. Desta forma, ocorre a redução da dimensão dos dados preservando somente as informações relevantes para a classificação.

Redes Neuronais Artificiais, Estatística, Algoritmos Genéticos, Lógica Fuzzy, Máquinas de Suporte Vetorial são alguns campos de conhecimento utilizados na área de Reconhecimento de Padrões (CARVALHO, 2005)(BANDYOPADHYAY, 2007). Outros métodos como teorema de Bayes, regra do vizinho mais próximo (k-NN) são também utilizados. Esta dissertação adotou o método de Reconhecimento de Padrões conhecido como "*Template Matching*" para a tarefa de classificação de criptogramas. Maiores detalhes serão informados no Capitulo 5 referente a metodologia de classificação.

2.1.2.2 TEMPLATE MATCHING

Neste método, a caracterização de padrões, a *priori* desconhecidos, faz-se por meio de comparações com um modelo previamente armazenado cujas características são usadas como parâmetro para a comparação. A figura 2.9 mostra um modelo clássico de classificação de dados utilizando o *Template Matching*.

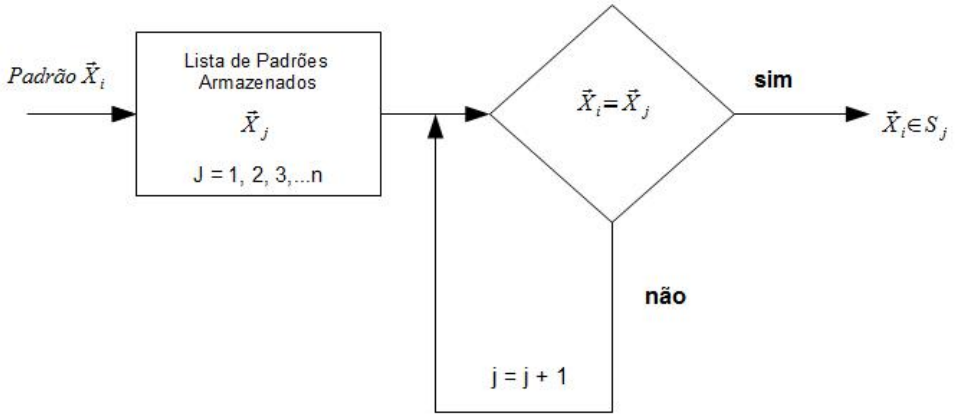


FIG. 2.9: Template Matching (PRADO, 2008).

Nesse método, um padrão desconhecido $\vec{X}_i$ é comparado com uma listagem de padrões previamente armazenados $\vec{X}_j$ $(j = 1, 2, 3, \ldots n)$. Se houver correlacionamento entre o vetor desconhecido e o conhecido durante a comparação, ocorre a classificação do parâmetro desconhecido.

2.1.2.3 APLICAÇÕES DO RECONHECIMENTO DE PADRÕES

Em diversos trabalhos e literaturas, de acordo com (BANDYOPADHYAY, 2007), são encontradas diversas aplicações do processo automático de Reconhecimento de Padrões em muitas áreas como:

- medicina: diagnoses médicas, análises de imagem, classificação de doenças;
- biologia computacional: identificação de genes, modelagem de proteínas;
- previsão e estudo de de recursos naturais: agricultura, geologia, meio-ambiente;
- biometria: reconhecimento de faces;
- veicular: automobilismo, controle de trens, aviões e barcos;

- defesa: reconhecimento automático de alvos;
- criminal: análise de impressões digitais e fotografias;
- industria: inspeção e controle de qualidade.

Entretanto, esta dissertação realizou pesquisas direcionadas na busca por trabalhos relacionados a aplicação de Algoritmos Genéticos no Reconhecimento de Padrões criptográficos. Convém ressaltar que o Reconhecimento de Padrões utilizado neste trabalho, por meio do Algoritmo Genético modelado, é uma ferramenta que pode auxiliar ou contribuir para uma fase preliminar da criptoanálise [13] [14]. (DILEEP, 2006) afirma que a identificação de métodos de criptografia é considerada como uma tarefa de Reconhecimento de Padrões.

Relembra-se que este trabalho tem como objetivo agrupar e classificar os criptogramas gerados pelos algoritmos criptográficos finalistas do AES, sem a necessidade de conhecer a quantidade e o tipo de algoritmos distintos em análise, pois a classificação de um determinado criptograma permite a identificação da cifra que foi utilizada para gerá-lo.

2.1.2.4 APLICAÇÃO DO ALGORITMO GENÉTICO EM CRIPTOLOGIA

Os Algoritmos Genéticos foram unitariamente utilizados em alguns trabalhos de criptoanálise com o objetivo de "quebrar" algumas cifras clássicas por meio da identificação das chaves. A Tabela 2.1 e 2.2 registram simplificadamente alguns trabalhos criptoanalíticos que utilizaram apenas o Algoritmo Genético. (DELMAN, 2004) criticou o uso dos Algoritmos Genéticos na criptoanálise e fez duas afirmações: "Nenhuma das cifras modernas, como o DES, AES, RSA, ou curva elíptica, foram usadas em uma abordagem criptoanalítica com algoritmo genético" e " As cifras que somente são consideradas, mesmo que remotamente, matematicamente difíceis são as que pertencem ao sistema *knapsack*, e mesmo elas não são verdadeiras referências. Esta ausência (resultados relevantes) faz com que muitos dos ataques com algoritmo genético tenham pouca importância no campo da criptoanálise".

[13] A criptoanálise trata da "quebra" de uma cifra para recuperar informações (MENEZES, 1996) (STALLINGS, 2008) (SINGH, 2008).

[14] Ataque criptoanalítico por "Só-texto-ilegível".

[1] As chaves foram encontradas com um elevado grau de acertos.

[2] Os resultados inconsistentes indicam que a cifra atacada não é a cifra de Vernam.

Tipo de Cifra	Permutação e Substituição	Vernam	Viginère
(CLARK, 1994)	chave encontrada [1]	X	X
(LIN, 1995)	X	Resultados inconsistentes [2]	X
(CLARK, 1997)	X	X	chave não encontrada

TAB. 2.1: Análise dos resultados de ataques criptoanalíticos realizados com Algoritmos Genéticos (DELMAN, 2004).

Tipo de Cifra	Substituição	Transposição	Knapsack
(JANSSEN, 1993)	chave não encontrada [3]	X	X
(MATTHEWS, 1993)	X	chave encontrada [4]	X
(SPILLMAN, 1993)	X	X	chave encontrada [5]
(YASEEN, 1999)	X	X	Resultados inconsistentes [6]
(GRUNDLINGH, 2002)	chave não encontrada	X	X

TAB. 2.2: Análise dos resultados de ataques criptoanalíticos realizados com Algoritmos Genéticos (DELMAN, 2004).

Como citado com a seção 2.1.2.3 , o Algoritmo Genético tem sido aplicado em diversas áreas de estudo. No caso específico da área criptológica, poucos trabalhos tem sido encontrados, conforme as Tabela 2.1 e 2.2. Entretanto, dois anos após as críticas de (DELMAN, 2004), (GARG, 2006) conseguiu, por meio do Algoritmo Genético, recuperar algumas chaves em criptogramas gerados por uma versão apenas simplificada do DES. Este resultado mostrou uma expectativa promissora da aplicação do Algoritmo Genético para a "quebra" de uma cifra de bloco. Todavia, até o presente momento, nenhum trabalho científico utilizando Algoritmos Genéticos apresentou resultado satisfatório no que tange a decriptografia não autorizada de alguma cifra de bloco pertencente ao grupo dos cinco algoritmos criptográficos finalistas do concurso do AES. Esta dissertação utiliza o Algoritmo Genético modelado como uma ferramenta auxiliar na criptoanálise. O tipo de ataque mais oneroso para um criptoanalista é o "Ataque por só-texto-ilegível". Neste tipo de ataque, por exemplo, o criptoanalista pode ter que testar todas as combinações de chaves em todos os tipos de cifras existentes em uso. Uma chave de 128 bits gera um universo de 2^{128} ou $3,4 \times 10^{38}$ chaves disponíveis. A identificação do tipo de cifra que foi usada para criptografar determinado arquivo acarreta em uma redução do esforço para este tipo de criptoanálise. Isto pode ser conseguido por meio do Algoritmo Genético.

[3] Cifra de substituição monoalfabética.

[4] 13.33% de acertos.

[5] O resultado foi considerado irrelevante em virtude da baixa complexidade da cifra.

[6] Os resultados inconsistentes em virtude da falta de informações dos parâmetros utilizados no ataque.

3 FASE DE *PRÉ-PROCESSAMENTO*

3.1 MODELAGEM VETORIAL DOS CRIPTOGRAMAS

A fase de *pré-processamento* é necessária para modelar os criptogramas em uma estrutura de dados que possa ser utilizada como dado de entrada no Algoritmo Genético modelado. Desta forma foi utilizado o modelo apresentado no trabalho de (CARVALHO, 2006) e (SOUZA, 2007), que considera os criptogramas que compõem uma coleção como um espaço de vetores de n-dimensões, onde n é o número de blocos binários no universo dos criptogramas. 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 frequência $f_{n,2}$ 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 por $\vec{C}_1 = (f_{1,1}, f_{2,1}, ..., f_{n,1})$. Mostra-se, na Figura 3.1, o processo de formação dos vetores dos criptogramas. 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.

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	$f_{n,1}$	$f_{n,2}$	$f_{n,3}$	$f_{n,4}$	.....	$f_{n,i}$

FIG. 3.1: Dicionário de blocos. $f_{n,i}$ é a frequência do n-ésimo bloco do i-ésimo criptograma da coleção de criptogramas

3.1.1 SIMILARIDADE ENTRE OS CRIPTOGRAMAS

Para avaliar o grau de associação (similaridade) entre o criptograma 1 e o criptograma 2, faz-se uma correlação entre os seus vetores $\vec{C}_1$ e $\vec{C}_2$. Esta correlação pode ser quantificada pelo *co-seno* do ângulo entre esses dois vetores na equação 3.1.

$$cos(\vec{C}_1, \vec{C}_2) = \frac{\sum_{i=0}^{n}(\vec{C}_1 * \vec{C}_2)}{\sqrt{\sum_{i=0}^{n}\vec{C}_1^2 * \sum_{i=0}^{n}\vec{C}_2^2}} \tag{3.1}$$

Assim, a modelagem de cada criptograma em um vetor de blocos binários quantifica a frequência com que estes blocos aparecem no criptograma e nos permite determinar o grau de associação entre os criptogramas. Este grau é uma medida de similaridade entre um par de vetores. Deste modo, calcula-se a similaridade entre todos os criptogramas de uma determinada coleção gerando uma da matriz de similaridades. A figura 3.2 ilustra a fase do *pré-processamento*.

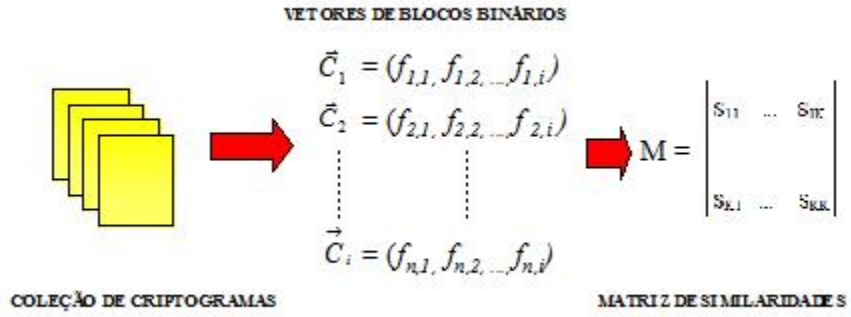


FIG. 3.2: Fase de *Pré-processamento.*

(CARVALHO, 2006) teve uma contribuição fundamental no processo de Reconhecimento de Padrões criptográficos, porque deu início a essa área de pesquisa no IME e contribuiu com a consolidação da fase de *pré-processamento* dos dados. (SOUZA, 2007) contribuiu com 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, confirmando o melhor desempenho da medida de similaridade *co-seno* dentro do contexto proposto. Por este motivo, a medida *co-seno* foi adotada como medida de similaridade no Algoritmo Genético modelado.

4 MODELAGEM DO ALGORITMO GENÉTICO

4.1 REPRESENTAÇÃO *CROMOSSOMIAL*

Um conjunto de grupos de criptogramas é representado por meio de uma matriz binária que será chamada de *cromossomo*. Esta matriz é visualizada na figura 4.1. 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. Este modelo binário de representação *cromossomial* está fundamentado em (GOLDSCHMIDT, 2005).

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

FIG. 4.1: Modelo representativo do *cromossomo* do Algoritmo Genético

4.2 DESCRIÇÃO SUCINTA DA DINÂMICA DO FUNCIONAMENTO DO ALGORITMO GENÉTICO MODELADO

A figura 4.2 ilustra o funcionamento dinâmico do Algoritmo Genético modelado. São gerados aleatoriamente 200 [15] *cromossomos* de acordo com o modelo *cromossomial* da figura 4.1. Um conjunto de *cromossomos* forma uma população. Convém ressaltar que o *cromossomo* [16] representa uma solução boa ou ruim para o Algoritmo Genético e que cada um dos *cromossomos* é um conjunto de grupos distintos que contém criptogramas. O objetivo do Algoritmo Genético é encontrar o melhor *cromossomo* da população e utilizá-lo para configurar um conjunto de grupos distintos com os criptogramas mais similares entre si.

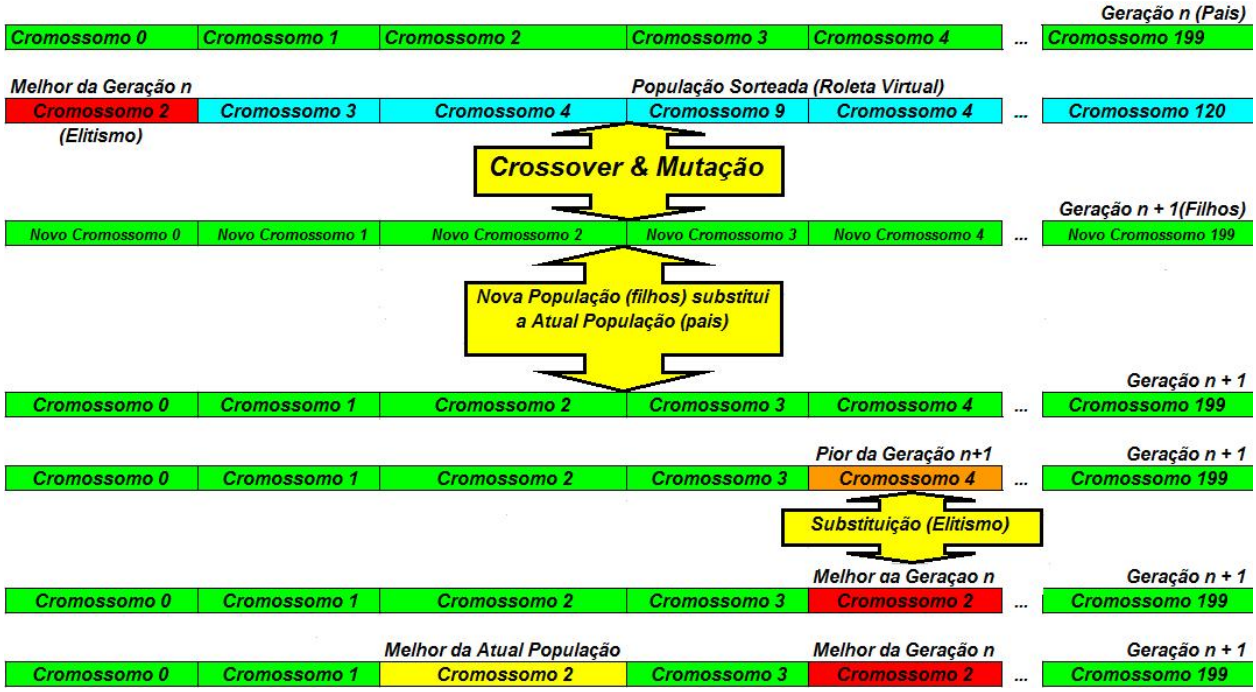


FIG. 4.2: Funcionamento dinâmico Algoritmo Genético modelado.

[15] Este valor de "200 *cromossomos*" foi arbitrado em virtude dos bons resultados encontrados nos ensaios realizados no capitulo 6 desta dissertação.

[16] Os termos ***cromossomo*** e **indivíduo** tem o mesmo significado na área de Algoritmos Genéticos.

O desempenho do Algoritmo Genético é dependente do tamanho da população. Se a população for pequena demais, será incapaz de achar uma boa solução em virtude do pouco do espaço para a "diversidade genética" [17]. Caso o Algoritmo Genético possua uma grande população, ele consumirá mais tempo e estará se aproximando de um algoritmo de busca exaustiva. Devido aos bons resultados encontrados nos ensaios realizados no capítulo 6, arbitrou-se o tamanho da população em 200 *cromossomos*.

Depois da geração dos primeiros 200 *cromossomos* que formam a primeira população da "geração n", cada um deles é avaliado pela função de avaliação *Calinski-Harabazs* (CH) cujo objetivo é determinar a qualidade do melhor *cromossomo*. Esta função será apresentada na seção 4.3. Após a avaliação de todos os *cromossomos* da primeira população da "geração n", o melhor *cromossomo* avaliado é selecionado para compor uma segunda população chamada de "população sorteada". Esta ação garante que, no "pior caso", o melhor *cromossomo* da primeira população da "geração n" participe da formação da "população sorteada". Este método é chamado de "elitismo". Depois, os outros 199 *cromossomos* da primeira população da "geração n" são submetidos a um sorteio para também formar, juntamente com o melhor *cromossomo* escolhido pelo "elitismo", a "população sorteada" de 200 *cromossomos*. Este sorteio dos *cromossomos* é realizado por meio de uma "roleta virtual" implementada na programação do Algoritmo Genético modelado.

Cabe ressaltar que no sorteio realizado pela "roleta virtual", existe a possibilidade de determinados *cromossomos*, com alta ou baixa qualidade, serem repetidos ou eliminados na composição da segunda população da "geração n". Deste modo, temos uma segunda população composta pelo melhor *cromossomo* da primeira população da "geração n" escolhido pelo método do "elitismo", concatenado com os *cromossomos* escolhidos de forma aleatória da primeira população. Estes 200 *cromossomos* da "população sorteada" formam então, uma segunda população da "geração n" que será submetida a transformações por meio de operadores genéticos. O objetivo dessas transformações é criar a primeira população da "geração $n + 1$". Destarte, a "população sorteada" é então submetida a operações de *crossover* e mutação. Estes tipos de operações genéticas serão apresentadas na subseção 4.2.1. Assim, temos os 200 *cromossomos* da "população sorteada" que são submetidos a 100 cruzamentos genéticos que acarreta em 200 novos *cromossomos* que

[17] A variedade genética são as diferentes "características genéticas" encontradas nos *cromossomos* de uma população. Vide subseção 3.2.1.3.

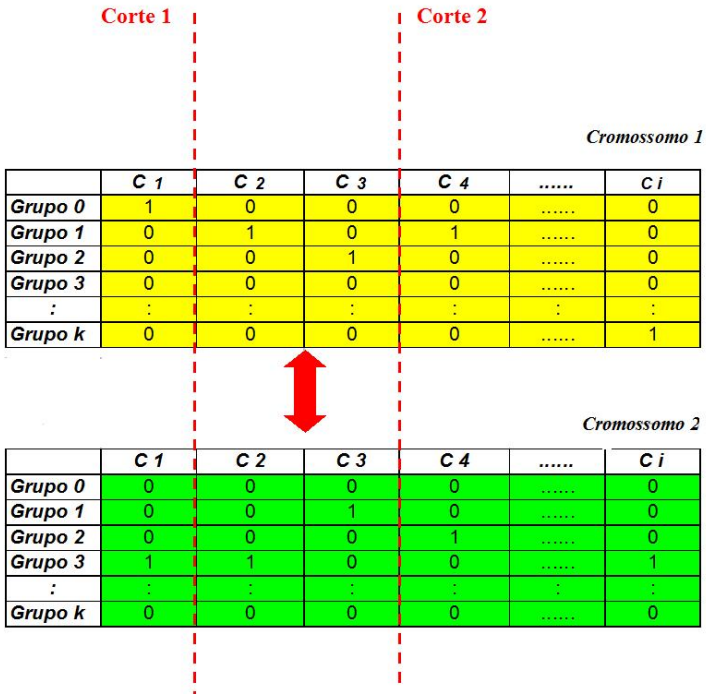
formarão a nova população de *cromossomos* ou primeira população da "geração $n+1$".

Ressalta-se que cada cruzamento genético da "população sorteada", irá gerar dois filhos (dois novos *cromossomos*). Para melhorar ainda mais a qualidade da primeira população da "geração $n+1$", após as operações genéticas, o seu pior *cromossomo* avaliado pela função CH é substituído pelo melhor *cromossomo* da primeira população da "geração n". Destarte, o método do "elitismo" é aplicado novamente, garantindo que o melhor *cromossomo* da primeira população da "geração n" não "morra" na primeira população da "geração $n+1$". Após todo este processo, a função CH é utilizada para selecionar o melhor *cromossomo* da primeira população da "geração $n+1$" que representa a melhor solução do problema.

4.2.1 OPERADORES GENÉTICOS

4.2.1.1 *CROSSOVERS*

O *crossover* é o cruzamento genético de dois *cromossomos*(pais) acarretando a geração de dois novos *cromossomos*(filhos). Em virtude dos bons resultados encontrados nos ensaios deste trabalho, a taxa do operador de *crossover* foi arbitrada em 95%. Isto significa que a probabilidade de ocorrência de um cruzamento genético entre dois *cromossomos* após a formação da "população sorteada" é elevada. Segundo (LINDEN, 2008), o operador de *crossover* tem recebido historicamente uma percentagem que varia de 60% a 95%. Além da taxa do *crossover*, para a realização do cruzamento genético, é necessário determinar o "ponto de corte". Este ponto constitui uma posição entre duas colunas da matriz binária (*cromossomo*) da figura 4.1. No Algoritmo Genético modelado, foi convencionado o esquema de *crossover* de "dois pontos de corte" aleatórios. Durante as operações de *crossover*, os criptogramas dinamicamente mudarão de grupos. Nas figuras 4.3 e 4.4, por exemplo, observa-se que os criptogramas C_2 e C_3 trocarão de grupos após o cruzamento genético de dois *cromossomos*. No *cromossomo* 1, os criptogramas C_2 e C_3 estão localizados nos Grupos 1 e 2, respectivamente. Do mesmo modo, no *cromossomo* 2, os criptogramas C_2 e C_3 pertencem aos Grupos 3 e 1, respectivamente. Assim, após o *crossover*, no novo cromossomo 1 (**filho 1**), os criptogramas C_2 e C_3 estão localizados nos Grupos 3 e 1 e, no novo cromossomo 2 (**filho 2**), os C_2 e C_3 pertencem aos Grupos 1 e 2, respectivamente. Destarte, temos dois novos *cromossomos* ou dois novos conjuntos de grupos de criptogramas.



	C 1	C 2	C 3	C 4	.....	C i
Grupo 0	1	0	0	0	.....	0
Grupo 1	0	1	0	1	.....	0
Grupo 2	0	0	1	0	.....	0
Grupo 3	0	0	0	0	.....	0
:	:	:	:	:	:	:
Grupo k	0	0	0	0	.....	1

	C 1	C 2	C 3	C 4	.....	C i
Grupo 0	0	0	0	0	.....	0
Grupo 1	0	0	1	0	.....	0
Grupo 2	0	0	0	1	.....	0
Grupo 3	1	1	0	0	.....	1
:	:	:	:	:	:	:
Grupo k	0	0	0	0	.....	0

FIG. 4.3: Dois pontos de corte aleatórios. Os *cromossomos* 1 e 2 são os pais que submetidos ao cruzamento genético formarão dois filhos.

Filho 1

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

Filho 2

	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

FIG. 4.4: 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".

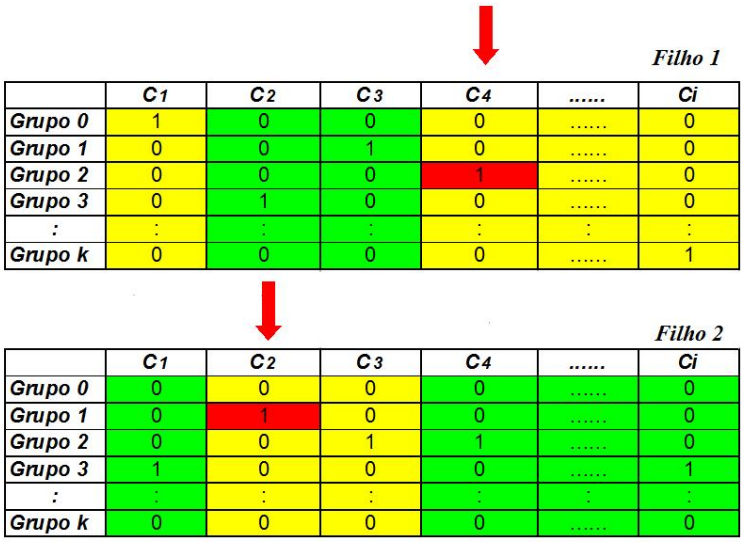
4.2.1.2 MUTAÇÃO

O operador de mutação por definição atua sobre um determinado *gene*, alterando-lhe o valor de forma aleatória. Isto significa que após a operação de *crossover*, um criptograma é escolhido aleatoriamente em cada novo *cromossomo* para mudar de grupo. Deste modo, no Algoritmo Genético modelado para cada *cromossomo* da primeira população da "geração $n+1$", um criptograma é escolhido aleatoriamente e os valores da sua correspondente coluna são zerados. Após 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 corresponde a afirmar que determinado criptograma não mudou de grupo. Para ilustrar esta situação específica, observa-se nas figuras 4.4 e 4.5 em que o criptograma C_2 (**filho 2**) não foi classificado em outro grupo, permanecendo no Grupo 2 após a mutação.

A taxa de mutação adotada no Algoritmo Genético modelado é de 1%. Isto significa que a probabilidade de ocorrer a mudança de grupo de um determinado criptograma é pequena em relação a taxa de *crossover*. Observa-se na figura 4.4 que o criptograma C_4 (**filho 1**) que antes pertencia ao Grupo 1, após a mutação, agora pertence ao Grupo 2 (Vide figura 4.5).

OS *CROMOSSOMOS* DE MENOR AVALIAÇÃO NÃO DEVEM SER DESCONSIDERADOS

Na Seção 4.2 foi mostrado que após a geração dos 200 *cromossomos* da primeira população da "geração n", é realizado um sorteio de 199 *cromossomos*. Os *cromossomos* (pais) com maior ou menor avaliação geram novos *cromossomos* (filhos). Existe a possibilidade, já que o sorteio é aleatório, que a "população sorteada" seja composta totalmente por *cromossomos* de baixa ou alta avaliação. No caso especifico de *cromossomos* de alta avaliação, o método do "elitismo" garante pelo menos a manutenção do *cromossomo* de maior avaliação. Entretanto, no caso de somente haver *cromossomos* de alta avaliação, o Algoritmo Genético poderá rapidamente sofrer uma "convergência genética". Relembra-se que os números binários de uma coluna da matriz binária do modelo "Representativo *cromossomial*" (vide Figura 4.1) é a informação de qual grupo pertence determinado

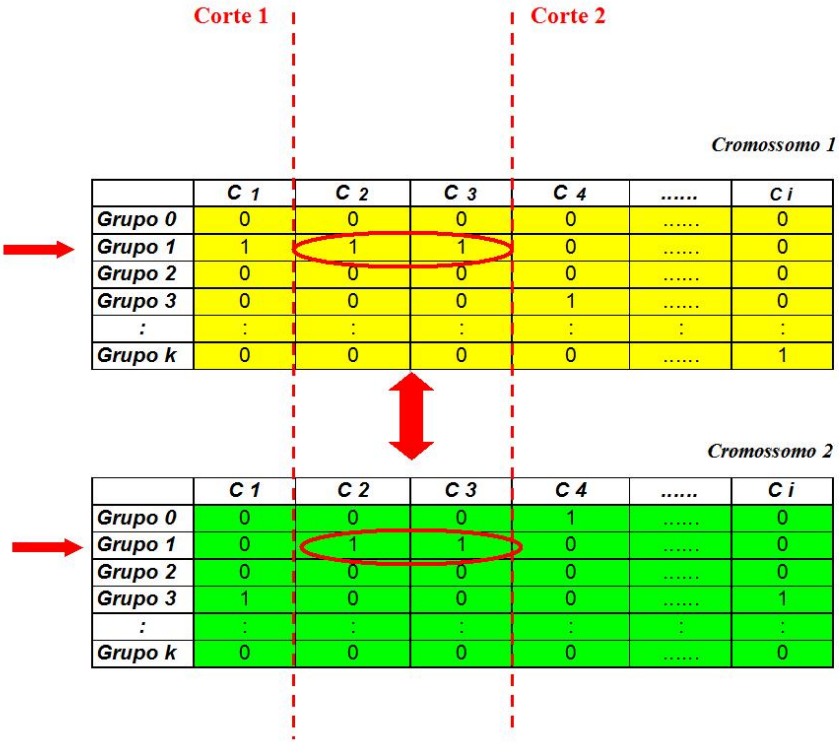


	C1	C2	C3	C4	.....	Ci
Grupo 0	1	0	0	0	.....	0
Grupo 1	0	0	1	0	.....	0
Grupo 2	0	0	0	1	.....	0
Grupo 3	0	1	0	0	.....	0
:	:	:	:	:	:	:
Grupo k	0	0	0	0	.....	1

	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

FIG. 4.5: Operação de mutação após a operação de *crossover*.

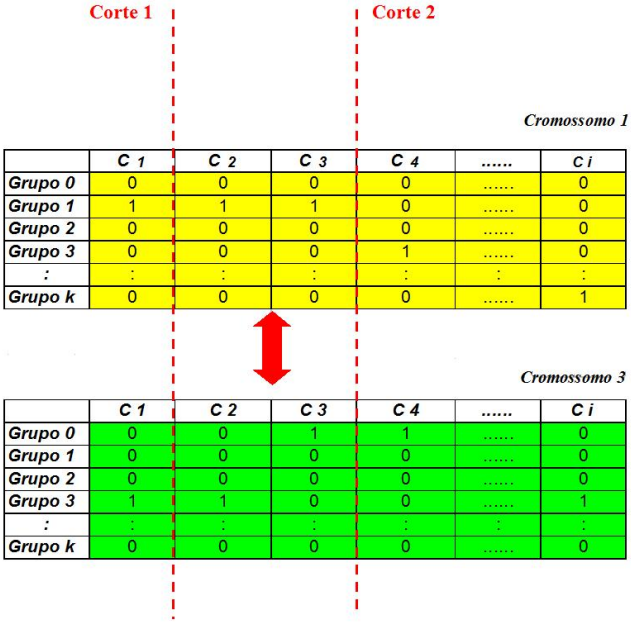
criptograma. Deste modo, cada coluna da matriz binária é uma "característica genética" do *cromossomo*. Na figura 4.6, por exemplo, considerando os *cromossomos* 1 e 2 com uma alta avaliação, observa-se que os criptogramas C_2 e C_3 pertencem ao mesmo Grupo 1. Neste caso específico, o *crossover* entre estes dois *cromossomos* não irá ocasionar nos criptogramas uma mudança de grupo. Assim, houve uma " convergência genética" no cruzamento genético entre os *cromossomos* 1 e 2. Na figura 4.7, por exemplo, considere que o *cromossomo* 3 possua uma baixa avaliação. Observa-se que o *cromossomo* 3 não detém as mesmas "características genéticas" dos *cromossomos* 1 e 2 da figura 4.6. Neste caso, um possível *crossover* entre o *cromossomo* 3 e o *cromossomo* 1, por exemplo, irá gerar novos *cromossomos*(filhos), acarretando uma mudança de grupo nos criptogramas C_2 e C_3.



	C 1	C 2	C 3	C 4	.....	C i
Grupo 0	0	0	0	0	.....	0
Grupo 1	1	1	1	0	.....	0
Grupo 2	0	0	0	0	.....	0
Grupo 3	0	0	0	1	.....	0
:	:	:	:	:	:	:
Grupo k	0	0	0	0	.....	1

	C 1	C 2	C 3	C 4	.....	C i
Grupo 0	0	0	0	1	.....	0
Grupo 1	0	1	1	0	.....	0
Grupo 2	0	0	0	0	.....	0
Grupo 3	1	0	0	0	.....	1
:	:	:	:	:	:	:
Grupo k	0	0	0	0	.....	0

FIG. 4.6: Convergência genética. Os *cromossomos* 1 e 2 possuem entre eles as mesmas "características genéticas" nos *genes* correspondentes aos criptogramas C_2 e C_3.



	C 1	C 2	C 3	C 4	.....	C i
Grupo 0	0	0	0	0	.....	0
Grupo 1	1	1	1	0	.....	0
Grupo 2	0	0	0	0	.....	0
Grupo 3	0	0	0	1	.....	0
:	:	:	:	:	:	:
Grupo k	0	0	0	0	.....	1

	C 1	C 2	C 3	C 4	.....	C i
Grupo 0	0	0	1	1	.....	0
Grupo 1	0	0	0	0	.....	0
Grupo 2	0	0	0	0	.....	0
Grupo 3	1	1	0	0	.....	1
:	:	:	:	:	:	:
Grupo k	0	0	0	0	.....	0

FIG. 4.7: O *cromossomo* 3 não possue as mesmas "características genéticas" do *cromossomo* 1.

PORQUE A MUTAÇÃO É NECESSÁRIA

Por definição, a "convergência genética" corresponde a uma população com baixa "diversidade genética", que por possuir *genes* [18] similares, não consegue evoluir. Desta maneira, mesmo após o cruzamento genético dos *cromossomos*, os criptogramas permanecem em seus grupos. Por esta razão, o operador de mutação é necessário para a garantir a existência de "diversidade genética" nos *cromossomos*. Por exemplo, observa-se na figura 4.8 que a existência de uma mutação, após o cruzamento genético ocorrido entre os *cromossomos* (pais) da figura 4.6, contribui para a continuidade da existência de "diversidade genética" na população de *cromossomos*. Assim, o criptograma C_3, após o *crossover* entre os *cromossomos* 1 e 2 na figura 4.6, é deslocado do Grupo 1 para o Grupo 3 em virtude da mutação.

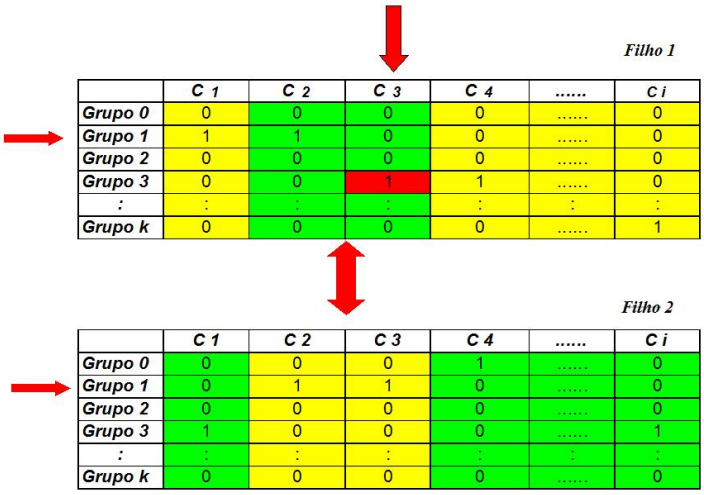


FIG. 4.8: Operação de mutação, após a operação de *crossover*, necessária para aumentar a "diversidade genética".

[18] Um *gene* de um *cromossomo* corresponde a coluna da matriz binária. Vide figura 4.1

PORQUE O OPERADOR DE MUTAÇÃO TEM UMA TAXA DE PROBABILIDADE MENOR DO QUE O OPERADOR DE *CROSSOVER*

Segundo (LINDEN, 2008), *"Não existe nenhum tipo de restrição em termos de associação das percentagens dos operadores. Qualquer um dos dois pode ter uma percentagem maior do que o outro. Entretanto, se a mutação for dominante, o seu Algoritmo Genético se parecerá muito com uma random walk* [19]*, e seus resultados serão equivalentes. Caso a probabilidade do operador de mutação for baixa demais, ela agirá de forma extremamente parcimoniosa e a população não terá diversidade depois de um certo número de gerações, estagnando bem rápido devido à convergência genética".*

Um outro fato para tentar justificar o baixo valor da probabilidade do operador de mutação é que sendo os Algoritmos Genéticos uma técnica de busca baseada numa metáfora do processo biológico de evolução natural, na natureza a mutação dos *genes* de um individuo é aplicada de forma menos frequente do que a reprodução sexuada (*crossover*).

A NECESSIDADE DOS OPERADORES GENÉTICOS

Nas seções e subseções anteriores, foram apresentados os operadores genéticos e descrito o seu funcionamento dinâmico no Algoritmo Genético modelado. Tanto as operações de *crossover* quanto as operações de mutação são necessárias para proporcionar a "evolução" da população de *cromossomos*. O *cromossomo* mais bem avaliado representa a melhor solução para o problema em análise. A "evolução" supracitada refere-se no Algoritmo Genético modelado à geração de mais conjuntos de grupos de criptogramas com o objetivo de encontrar como solução ótima um conjunto de grupos de criptogramas mais similares.

Por analogia, as operações de *crossover* e mutação correspondem à ações de permutação e substituição, respectivamente. A permutação é realizada entre as colunas (*genes*) das matrizes binárias (*cromossomos*) e a substituição é executada em uma única coluna (*gene*) de uma determinada matriz binária, após a operação de permutação. Estas permutações e substituições do Algoritmo Genético modelado são responsáveis em produzir deslocamentos de criptogramas de um determinado grupo para outro.

[19] "Random Walk" é um tipo de algoritmo de busca totalmente aleatória. Contrariamente, o Algoritmo Genético é uma técnica de busca "direcionada" em virtude das "informações" fornecidas pela população (*cromossomos*) corrente. Estas "informações" são o resultado das ações da função de avaliação e do cruzamento genético.

As técnicas de substituições e permutações de blocos binários são a base da criptografia do AES, e da maioria dos outros sistemas de cifrantes de blocos simétricos iterativos conhecidos (LAMBERT, 2004). Estas técnicas são utilizadas nas operações genéticas nas matrizes binárias (*cromossomos*) do Algoritmo Genético modelado para a detecção de padrões nas cifras geradas pelos algoritmos cifrantes.

4.3 FUNÇÃO DE AVALIAÇÃO

De acordo com (GOLDSCHMIDT, 2005), as funções de avaliação mais usuais para problemas envolvendo a tarefa de agrupamento são:

a) *Minimização do traço (W)*

$$W = \sum_{i=1}^{k} \sum_{j=1}^{n_k} \parallel x_{ij} - z_i \parallel^2 \tag{4.1}$$

,onde:

- k = Número de grupos ;
- n_i = Número de criptogramas do grupo i;
- z_i = Centro geométrico (centróide) de um grupo de criptogramas;
- x_{ij} = j- ésimo criptograma do grupo i.

A função de *Minimização do Traço (W)* é o somatório da distância vetorial entre cada criptograma de um determinado grupo em relação ao centróide [20] deste mesmo grupo. Quanto **menor** o valor do somatório de W, melhor será a homogeneidade interna de um grupo. É uma medida de dispersão interna dentro de um grupo de criptogramas.

As figuras 4.9 e 4.10 exemplificam um ambiente *bi-dimensional* que ilustra o comportamento de dois conjuntos de criptogramas analisado pela função de *Minimização de Traço (W)*. A figura 4.9 ilustra um grupo formado por cinco criptogramas C_1, C_2, C_3, C_4 e C_5. Cada criptograma está a uma determinada distância do centróide deste grupo e a função de *Minimização de Traço* possui um valor W_1. Na figura 4.10, os criptogramas C_6, C_7, C_8, C_9 e C_{10} estão a uma distância relativamente menor do centróide do seu grupo em comparação aos criptogramas da figura 4.9 e o valor da função *Minimização de Traço* deste segundo grupo é W_2. Deste modo, $W_2 < W_1$.

[20] A definição e o cálculo do centróide no Algoritmo Genético modelado podem ser vistas na seção 4.4

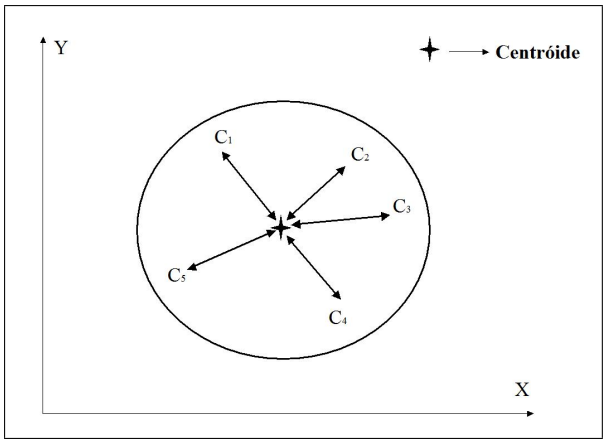


FIG. 4.9: Quanto menor o valor do somatório da distância vetorial W, mais homogêneo é o grupo formado.

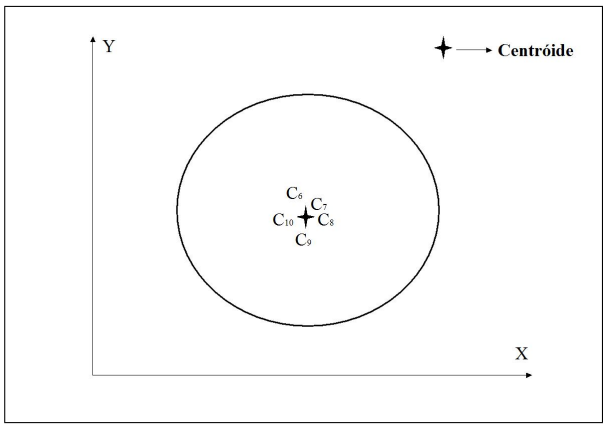


FIG. 4.10: Grupo homogêneo com criptogramas muito similares e muito próximos do centróide.

b) *Maximização do traço* ($\frac{B}{W}$);

$$\frac{B}{W} = \frac{\sum_{i=1}^{k} n_i \parallel z_i - z \parallel^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \parallel x_{ij} - z_i \parallel^2} \tag{4.2}$$

, onde:

- z = Centro geométrico (centróide) do conjunto de todos os criptogramas;

A função B é uma a medida de dispersão externa entre de os grupos de criptogramas. As figuras 4.11 e 4.12 mostram o comportamento de um conjunto de criptogramas analisado pela função de *Maximização de Traço* (BW^{-1}). Este conjunto tem dois subconjuntos distintos onde cada um é formado por cinco criptogramas similares. A figura 4.12 ilustra uma situação onde temos grupos bem definidos.

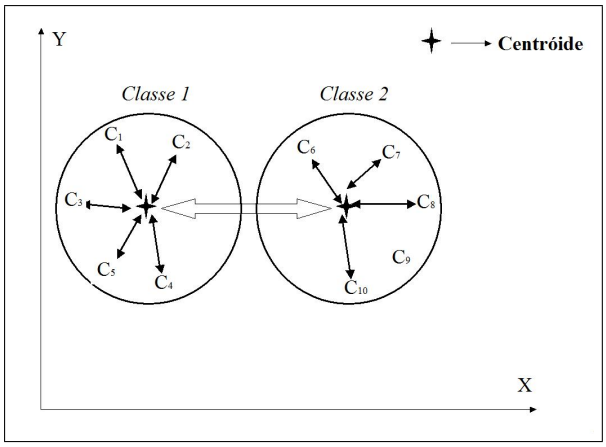


FIG. 4.11: Quanto menor o valor de B, mais próximos serão os centróides.

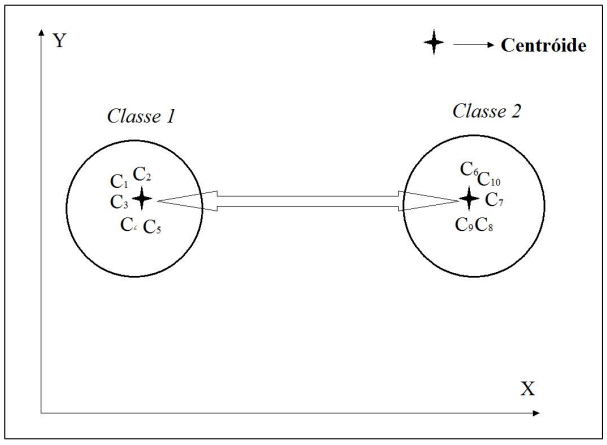


FIG. 4.12: Quanto maior o valor de B, mais distântes serão os centróides.

DISCUSSÃO

Ambas as funções apresentadas W e B, implementadas no Algoritmo Genético, realizam o agrupamento de uma coleção de criptogramas. Entretanto, há uma restrição correlacionada a estas funções no que diz respeito a quantidade de grupos gerados. Para a realização de um agrupamento (*clustering*), o valor do número de grupos torna-se um parâmetro indispensável de entrada no algoritmo modelado. É compreensível esta limitação pelo fato de (JAIN, 1999) nos mostrar que W é equivalente a função "erro-quadrático" utilizada no algoritmo *k-means*. Esta restrição de acordo com (HALKIDI, 2001), obriga o usuário especificar o número de grupos que se deve gerar em uma massa de dados. Para erradicar esta restrição, foi inserido no Algoritmo Genético modelado o índice $Calinski-Harabasz$ (CH) procedente de (CALINSKI, 1974). Assim, o agrupamento foi executado sem a necessidade de informar o número exato de grupos para o algoritmo modelado. (BANDYOPADHYAY, 2002) e (CONCI, 2008) utilizaram este índice em outros trabalhos de agrupamento automático com o algoritmo *k-means* em variados conjuntos de dados não correlacionados com criptogramas.

c)Índice *Calinski−Harabasz* (CH)

$$CH = \frac{B(k-1)}{W(n-k)} = \frac{\sum_{i=1}^{k} n_i \parallel z_i - z \parallel^2 (k-1)}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \parallel x_{ij} - z_i \parallel^2 (n-k)} \tag{4.3}$$

Este índice é a função de *Maximização de traço (B/W)* multiplicada pelo fator (k-1)/(n-k) que é similar ao termo de Hartigan (n-k-1) procedente de (HARTIGAN, 1975). Os termos multiplicadores tanto do índice CH quanto de Hartigan, de acordo com (BASSAB, 1990), são os graus de liberdade aplicados na função estatística. (LI, 2008) considera o termo de Hartigan como um fator de penalidade para a correção de um número grande de grupos.

Na próxima subseção é mostrado graficamente a diferença entre as funções de avaliação. Esta diferença foi importante para a automatização do processo de agrupamento entre os criptogramas.

4.3.1 COMPARAÇÃO GRÁFICA ENTRE AS FUNÇÕES DE AVALIAÇÃO

As seguintes subseções tem como objetivo mostrar o comportamento gráfico de cada tipo de função apresentada neste trabalho. Os gráficos são *bi-dimensionais*. O eixo das abcissas corresponde ao número k de grupos e o eixo das ordenadas corresponde ao valor encontrado pela função de avaliação para determinado k. Relembra-se que cada *cromossomo* do Algoritmo Genético modelado representa um conjunto de grupos de criptogramas. Para gerar os gráficos abaixo, foi utilizado um Banco de Dados (BD) formado por um total de 105 criptogramas. Neste BD temos cinco grupos distintos de criptogramas. Cada grupo é formado por 21 criptogramas mais similares entre si. Detalhes técnicos sobre o tipo de algoritmo cifrante usado ou chave serão mostrados somente em outros experimentos no capítulo 6.

4.3.1.1 *MINIMIZAÇÃO DO TRAÇO (W)*

Observando o gráfico da figura 4.13, percebe-se que a função de *Minimização de traço* apresenta uma curva gráfica crescente, proporcional aos valores de k. Relembra-se que a coleção de criptogramas analisada é formada por cinco grupos distintos. Neste gráfico, não há nenhuma indicação que o valor para k igual a 5 é a resposta correta para o agrupamento dos criptogramas.

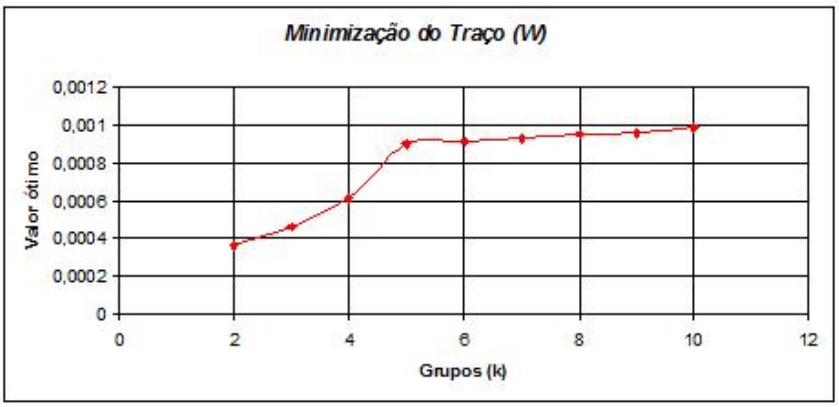


FIG. 4.13: Agrupamento de criptogramas com o número k de grupos variando de 2 a 10.

4.3.1.2 *MAXIMIZAÇÃO DO TRAÇO* (BW^{-1})

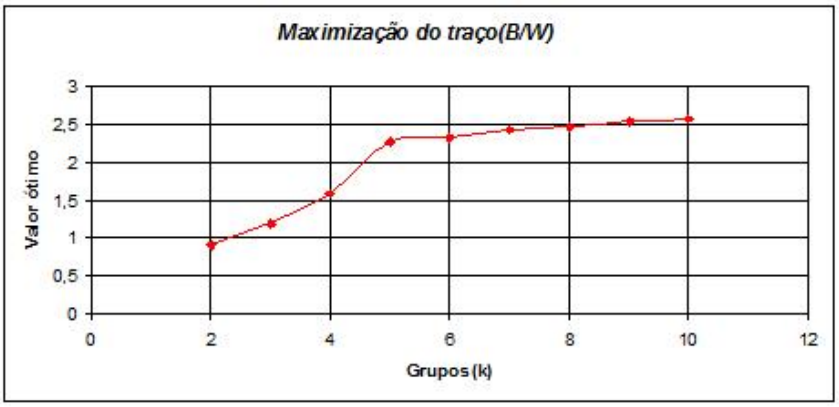


FIG. 4.14: Agrupamento de criptogramas com o número k de grupos variando de 2 a 10.

Analisando o gráfico da figura 4.14, observa-se uma representação gráfica semelhante ao da figura 4.13. Assim, não há nenhuma indicação para o valor correto de grupos.

4.3.1.3 ÍNDICE *CALINSKI−HARABASZ*(CH)

Na figura 4.15 observa-se que o gráfico gerado pela função *Calinski-Harabasz* apresenta um "joelho" na curva que também corresponde ao seu ponto máximo global. É neste ponto que está sendo observado o valor para k igual cinco. Este valor é o número correto de grupos na nossa coleção de criptogramas analisada.

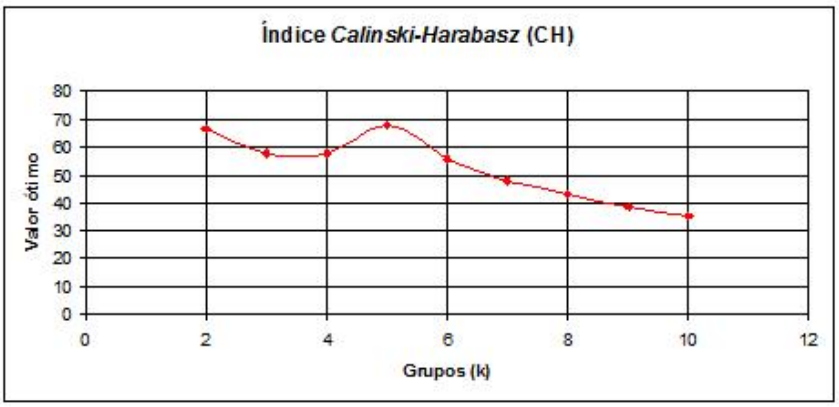


FIG. 4.15: Agrupamento de criptogramas com o número k de grupos variando de 2 a 10.

DISCUSSÃO

O índice *Calinski-Harabasz* foi a função de avaliação escolhida para o Algoritmo Genético modelado para a realização de agrupamentos de criptogramas. O seu resultado gráfico mostra o valor correto do número de grupos em uma amostra formada por criptogramas diferentes. O "joelho" observado na curva gráfica, segundo (CALINSKI, 1974), sugere o ponto que indica o valor correto de grupos formados em uma amostra de dados. Será visto, nos ensaios realizados no capitulo 6, que o "joelho" do gráfico que indica o número correto de grupos de criptogramas algumas vezes não coincide com o valor máximo global do próprio gráfico. A propriedade gráfica indicativa do número correto de grupos da função CH proporcionou ao Algoritmo Genético modelado uma automatização no agrupamento de criptogramas similares.

4.4 FUNCIONAMENTO DINÂMICO DA FUNÇÃO DE AVALIAÇÃO

Relembra-se que o Algoritmo Genético modelado gera *cromossomos* que correspondem a conjuntos de grupos de criptogramas. Apenas o *cromossomo* com a melhor avaliação representa a melhor solução do problema. Assim, o melhor *cromossomo* avaliado pela função *Calinski-Harabasz* (CH), é aquele que corresponde a um conjunto de grupos distintos formados por criptogramas mais similares entre si. Para realizar esta avaliação, a função CH utiliza como dados de entrada: uma matriz de similaridades de criptogramas e as informações (*genes*) de um *cromossomo*. A figura 4.16 ilustra a relação da função CH com os seus parâmetros de entrada.

Para mostrar o funcionamento da figura 4.16, considere, por exemplo, um conjunto

FIG. 4.16: Descrição sucinta do funcionamento dinâmico da função de avaliação.

de 5 criptogramas (C_1,C_2, C_3, C_4, e C_5. Este conjunto de criptogramas foi submetido a fase de *pré-processamento* [21] e gerou uma matriz de similaridades ilustrada na figura 4.17. Uma representação *cromossomial* [22] deste criptograma é mostrado na figura 4.18.

	C_1	C_2	C_3	C_4	C_5
C_1	1	$a_{1,2}$	$a_{1,3}$	$a_{1,4}$	$a_{1,5}$
C_2	$a_{2,1}$	1	$a_{2,3}$	$a_{2,4}$	$a_{2,5}$
C_3	$a_{3,1}$	$a_{3,2}$	1	$a_{3,4}$	$a_{3,5}$
C_4	$a_{4,1}$	$a_{4,2}$	$a_{4,3}$	1	$a_{4,5}$
C_5	$a_{5,1}$	$a_{5,2}$	$a_{5,3}$	$a_{5,4}$	1

FIG. 4.17: Matriz de similaridades dos criptogramas C_1, C_2, C_3, C_4 e C_5.

	C_1	C_2	C_3	C_4	C_5
Grupo 1	1	1	1	0	0
Grupo 2	0	0	0	1	1

FIG. 4.18: Representação *cromossomial* dos criptogramas C_1, C_2, C_3, C_4 e C_5.

A função CH realiza a avaliação do *cromossomo* da figura 4.18. A informação genética deste *cromossomo* é que os criptogramas C_1,C_2 e C_3 pertencem ao grupo 1 e os criptogramas C_4 e C_5 pertencem ao grupo 2. Após isto, O Algoritmo Genético modelado considera cada criptograma como um "vetor de similaridades" de todos os criptogramas do conjunto. Então, por meio da matriz de similaridades da figura 4.17, os criptogramas C_1, C_2, C_3, C_4 e C_5 são modelados vetorialmente da seguinte forma:

$$\vec{C_1} = (1, a_{1,2}, a_{1,3}, a_{1,4}, a_{1,5});$$

$$\vec{C_2} = (a_{2,1}, 1, a_{2,3}, a_{2,4}, a_{2,5});$$

[21] Vide capitulo 3, figura 3.2.
[22] Vide capitulo 4, figura 4.1.

$$\vec{C_3} = (a_{3,1},\ a_{3,2},\ 1,\ a_{3,4},\ a_{3,5});$$

$$\vec{C_4} = (a_{4,1},\ a_{4,2},\ a_{4,3},\ 1,\ a_{4,5});$$

$$\vec{C_5} = (a_{5,1},\ a_{5,2},\ a_{5,3},\ a_{5,4},\ 1).$$

CÁLCULO DO CENTRÓIDE

Assim, após a modelagem de cada criptograma em um "vetor de similaridades", o Algoritmo Genético modelado, com as informações genéticas do *cromossomo* da figura 4.18, calculará o centróide dos grupos 1 e 2.

Segundo (GOLDSCHMIDT, 2005) e (FREI, 2006), o centróide é a média aritmética das distâncias de cada objeto para os demais. Quanto mais similares forem os objetos, menor é a distância entre eles. No Algoritmo Genético modelado, estes objetos são os criptogramas que estão sendo analisados. Conforme visto no capitulo 3, os criptogramas são modelados em vetores de blocos binários e a relação entre cada vetor é medida pelo *co-seno* [23] do ângulo entre eles. Todas estas relações entre os vetores dos criptogramas gera uma matriz de similaridades. Quanto menor o ângulo entre dois vetores, maior é o valor do *co-seno* e menor é a distância [24] entre eles no espaço n-dimensional.

Neste trabalho, o centróide é uma grandeza vetorial calculada pela Raiz Média Quadrática (RMQ) dos "vetores de similaridades" dos criptogramas. A razão para o uso da métrica RMQ foi em virtude dos melhores resultados [25] apresentados na tarefa de agrupamento para o Algoritmo Genético modelado. O cálculo do vetor centróide ($\vec{z}$) é ilustrado na equação 4.4:

$$\vec{z} = \sqrt{\frac{\vec{C_1}^{\,2} + \vec{C_2}^{\,2} + \vec{C_3}^{\,2} + \cdots + \vec{C_n}^{\,2}}{n}} \tag{4.4}$$

, onde:

- $\vec{C_n}$ = "Vetor de similaridade" que representa o n-ésimo criptograma de um grupo;
- n = Número de "vetores de similaridades" de criptogramas do grupo.

[23] Vide capitulo 3, equação 3.1.

[24] Vide seção 4.3, figuras 4.9 e 4.10.

[25] Inicialmente, para o cálculo do centróide foi utilizado a métrica da média aritmética simples que acarretou em resultados insatisfatórios no agrupamento.

Deste modo, analisando o *cromossomo* da figura 4.18, o centróide do grupo 1, $\vec{z_{G1}}$, é calculado pela Raiz Média Quadrática dos "vetores de similaridades" $\vec{C_1}$, $\vec{C_2}$ e $\vec{C_3}$. Então:

$$\vec{z_{G1}} = \sqrt{\frac{\vec{C_1}^2+\vec{C_2}^2+\vec{C_3}^2}{3}} \therefore$$

$$z_{1,1} = \sqrt{\frac{(1)^2 + (a_{2,1})^2 + (a_{3,1})^2}{3}} \ ;$$

$$z_{1,2} = \sqrt{\frac{(a_{1,2})^2 + (1)^2 + (a_{3,2})^2}{3}} \ ;$$

$$z_{1,3} = \sqrt{\frac{(a_{1,3})^2 + (a_{2,3})^2 + (1)^2}{3}}.$$

$$z_{1,4} = \sqrt{\frac{(a_{1,4})^2 + (a_{2,4})^2 + (a_{3,4})^2}{3}} \ ;$$

$$z_{1,5} = \sqrt{\frac{(a_{1,5})^2 + (a_{2,5})^2 + (a_{3,5})^2}{3}}.$$

Assim, o vetor do centróide do grupo 1 é dado por $\vec{z_{G1}} = (z_{1,1}, z_{1,2}, z_{1,3}, z_{1,4}, z_{1,5})$. Analogamente, é calculado o vetor do centróide do grupo 2, $\vec{z_{G2}}$:

$$\vec{z_{G2}} = \sqrt{\frac{\vec{C_4}^2+\vec{C_5}^2}{2}} \therefore$$

$$z_{2,1} = \sqrt{\frac{(a_{4,1})^2 + (a_{5,1})^2}{2}} \ ;$$

$$z_{2,2} = \sqrt{\frac{(a_{4,2})^2 + (a_{5,2})^2}{2}} \ ;$$

$$z_{2,3} = \sqrt{\frac{(a_{4,3})^2 + (a_{5,3})^2}{2}};$$

$$z_{2,4} = \sqrt{\frac{(1)^2 + (a_{5,4})^2}{2}} \ ;$$

$$z_{2,5} = \sqrt{\frac{(a_{4,5})^2 + (1)^2}{2}}.$$

Deste modo, o vetor do centróide do grupo 2 é igual a $\vec{z_{G2}} = (z_{2,1}, z_{2,2}, z_{2,3}, z_{2,4}, z_{2,5})$.

Após o cálculo dos vetores que representam o centróide de cada grupo, a função de avaliação *Calinski-Harabasz* (CH) é aplicada. Relembra-se que o índice CH é a função de *Maximização de Traço* (BW^{-1}) multiplicada por um fator de penalidade [26].

[26]Vide capitulo 4, equação 4.3.

CÁLCULO DA MEDIDA DE DISPERSÃO INTERNA (W)

Aplicando a equação (4.1) para o grupo 1, então:

Grupo 1 - Critptogramas C_1, C_2 e C_3

$$W_1 = \|\vec{C_1} - \vec{z_{G1}}\|^2 \therefore$$

$$W_1 = ((1 - z_{1,1})^2 + (a_{1,2} - z_{1,2})^2 + (a_{1,3} - z_{1,3})^2 + (a_{1,4} - z_{1,4})^2 + (a_{1,5} - z_{1,5})^2);$$

$$W_2 = \|\vec{C_2} - \vec{z_{G1}}\|^2 \therefore$$

$$W_2 = ((a_{2,1} - z_{1,1})^2 + (1 - z_{1,2})^2 + (a_{2,3} - z_{1,3})^2 + (a_{2,4} - z_{1,4})^2 + (a_{2,5} - z_{1,5})^2);$$

$$W_3 = \|\vec{C_3} - \vec{z_{G1}}\|^2 \therefore$$

$$W_3 = ((a_{3,1} - z_{1,1})^2 + (a_{3,2} - z_{1,2})^2 + (1 - z_{1,3})^2 + (a_{3,4} - z_{1,4})^2 + (a_{3,5} - z_{1,5})^2);$$

Assim, a medida W_{G1} de dispersão interna do Grupo 1, formado pelos criptogramas C_1, C_2 e C_3, é calculada pelo somatório:

$$W_{G1} = W_1 + W_2 + W_3$$

Da mesma forma, os cálculos são repetidos para o grupo 2, então:

Grupo 2 - Criptogramas C_4 e C_5

$$W_4 = \|\vec{C_4} - \vec{z_{G2}}\|^2 \therefore$$

$$W_4 = ((a_{4,1} - z_{2,1})^2 + (a_{4,2} - z_{2,2})^2 + (a_{4,3} - z_{2,3})^2 + (1 - z_{2,4})^2 + (a_{4,5} - z_{2,5})^2);$$

$$W_5 = \|\vec{C_5} - \vec{z_{G2}}\|^2 \therefore$$

$$W_5 = ((a_{5,1} - z_{2,1})^2 + (a_{5,2} - z_{2,2})^2 + (a_{5,3} - z_{2,3})^2 + (a_{5,4} - z_{2,4})^2 + (1 - z_{2,5})^2);$$

Assim, a medida W_{G2} de dispersão interna do Grupo 2, formada pelos criptogramas C_4 e C_5, é também calculada pelo somatório:

$$W_{G2} = W_4 + W_5.$$

Destarte, calculado as medidas de dispersão interna de cada grupo, é realizado o somatório entre W_{G1} e W_{G2} com o objetivo de encontrar a medida de dispersão global interna W, indicada pelo *cromossomo* da figura 4.18. Então:

$$\mathrm{W} = W_{G1} + W_{G2}.$$

CÁLCULO DA MEDIDA DE DISPERSÃO EXTERNA (B)

Antes de aplicar a equação (4.2) para calcular a medida de dispersão externa dos Grupos 1 e 2, é necessário calcular o vetor ($\vec{z}$) do centróide de todo o conjunto dos criptogramas. A métrica RMQ é utilizada nos "vetores de similaridades" $\vec{C_1}$, $\vec{C_2}$, $\vec{C_3}$, $\vec{C_4}$ e $\vec{C_5}$. Então:

$$\vec{z} = \sqrt{\frac{\vec{C_1}+\vec{C_2}+\vec{C_3}+\vec{C_4}+\vec{C_5}}{5}}$$

, então:

$\vec{z} = (z_1, z_2, z_3, z_4, z_5) \longrightarrow$ Vetor que representa o ponto médio global (centróide) do conjunto de criptogramas.

Grupo 1 - Criptogramas C_1, C_2 e C_3

Aplicando a medida de dispersão externa B, numerador da equação (4.2), no Grupo 1:

$B_{G1} = n_{G1} * \|\vec{z_{G1}} - \vec{z}\|^2$, onde:

$n_{G1} \rightarrow$ O número de criptogramas do Grupo 1.

Grupo 2 - Criptogramas C_4 e C_5

Analogamente, para o Grupo 2 tem-se:

$$B_{G2} = n_{G2} * \|\vec{z_{G2}} - \vec{z}\|^2, \text{ onde:}$$

$n_{G2} \rightarrow$ O número de criptogramas do Grupo 2.

Assim, as medidas de dispersão externa de cada grupo foram calculadas. Após isto, é realizado o somatório entre B_{G1} e B_{G2} com o objetivo de encontrar a medida de dispersão global externa B. Então:

$$\text{B} = B_{G1} + B_{G2}.$$

CÁLCULO DA FUNÇÃO DE AVALIAÇÃO *CALINSKI-HARABASZ* (CH)

Após o cálculo da medidas de dispersão B e W dos grupos 1 e 2, aplica-se o índice CH com o seu fator de penalidade visto na seção 4.3. Então:

$$\text{CH} = \frac{B * (k-1)}{W * (n-k)} = \frac{B * (2-1)}{W * (5-2)} = \frac{B}{3 * W}, \text{ onde:}$$

$n \rightarrow$ Número total de criptogramas no conjunto analisado.

4.4.1 DISCUSSÃO

Para o funcionamento da função de avaliação no Algoritmo Genético modelado, percebe-se que os criptogramas foram submetidos a duas modelagens (vetorizações). A primeira modelagem refere-se a representação de cada criptograma em um vetor de blocos binários com o objetivo de gerar uma matriz de similaridades entre os criptogramas do conjunto total analisado. A segunda modelagem é a representação de cada criptograma em um "vetor de similaridades" com o objetivo de encontrar os centróides dos grupos e do conjunto total de criptogramas. Desta forma, a figura 4.19 ilustra sucintamente as duas modelagens aplicadas ao conjunto de criptogramas.

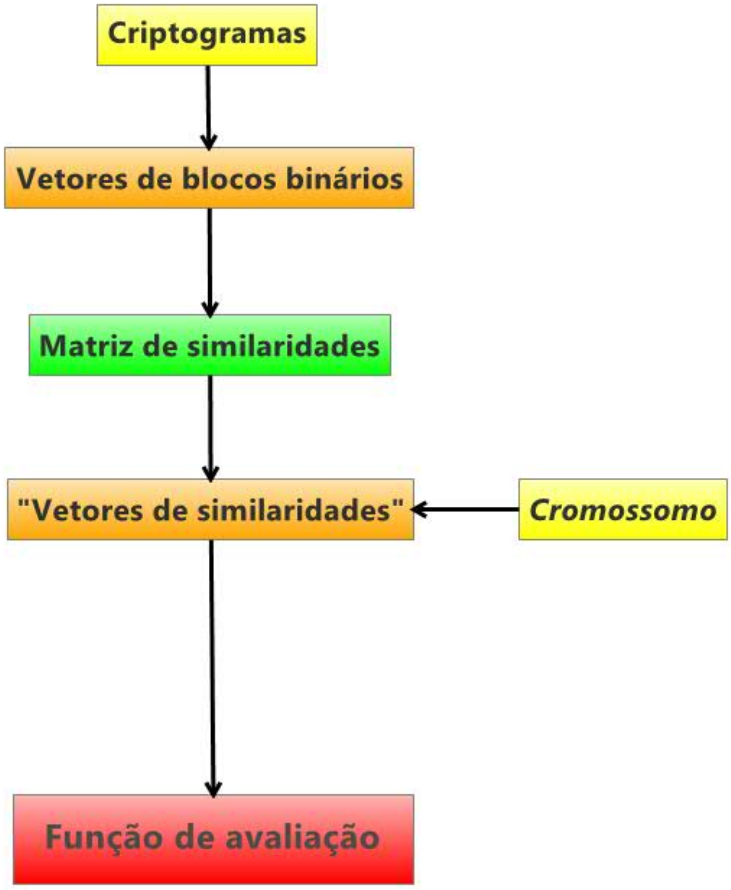


FIG. 4.19: Vetorização dos criptogramas para entrada na função de avaliação.

4.5 PARÂMETROS DE ENTRADA

A tabela 4.1 contém os parâmetros (*ad hoc*) utilizados no Algoritmo Genético modelado com o quais foram encontrados os melhores resultados.

Parâmetro	Valor
Tamanho da população	200
Número de gerações	3000
Taxa de *crossover*	0.95
Taxa de mutação	0.01

TAB. 4.1: Parâmetros específicos de entrada do Algoritmo Genético modelado no conjunto de criptogramas analisados.

Os parâmentros da tabela 1 foram utilizados em uma máquina com processador Core 2 Duo 1.8 GHz. O tempo de processamento do Algoritmo Genético, para cada experimento realizado no Capitulo 6, foi de 1 hora e 36 minutos. Entretanto, para uma máquina com processador Core 2 Duo 2.6 GHz, o tempo de processamento foi reduzido para 24 minutos.

4.6 MÉTRICAS UTILIZADAS PARA A AVALIAÇÃO DO AGRUPAMENTO

Para avaliar os resultados do agrupamento, foram utilizadas as métricas de revocação (*recall*) e precisão (*precision*). A revocação (r) é a razão entre a quantidade de criptogramas corretamente agrupados pelo sistema (cs) e a quantidade esperada de criptogramas agrupados (c). A precisão (p) é a proporção entre a quantidade de criptogramas corretamente agrupados pelo sistema (cs) e o total de criptogramas agrupados pelo sistema (na). Deste modo, considerando n grupos existentes, os valores de revocação e precisão são dados por:

$$r = \frac{1}{n}\sum_{i=1}^{n}(cs_i/c_i) \text{ e } p = \frac{1}{n}\sum_{i=1}^{n}(cs_i/a_i), \text{ onde:}$$

- cs_i = Número de criptogramas corretamente agrupados no grupo i;
- c_i = Número esperado de criptogramas no grupo i;
- a_i = Número de criptogramas agrupados no grupo i.

A figura 4.20 ilustra as métricas apresentadas que foram também utilizadas por (CARVALHO, 2006) e (SOUZA, 2007).

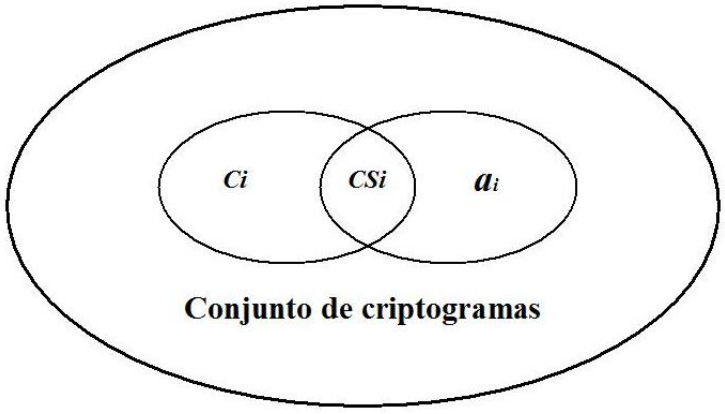


FIG. 4.20: Medidas de precisão e revocação.

Para melhor compreensão da aplicação destas métricas, considere uma coleção de 20 criptogramas a qual possui 5, 8 e 7 textos cifrados pelo AES, RC6 e MARS, respectivamente. Esta coleção é submetida a 3 operações distintas de agrupamento e fornece três possíveis exemplos:

PRIMEIRO EXEMPLO - 3 GRUPOS FORMADOS

No primeiro exemplo, os 3 grupos formados possuem as configurações ilustradas pela figura 4.21.

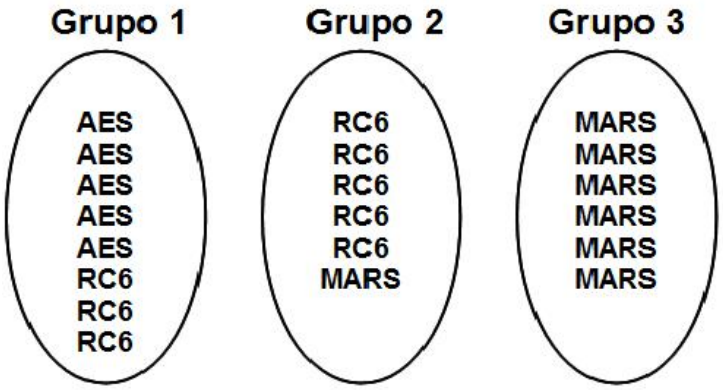


FIG. 4.21: Agrupamento com 3 grupos formados. Precisão = 0.82 e Revocação = 0.75.

Aplicando as fórmulas matemáticas de precisão(p) e revocação(r), tem-se:

$$p = \frac{1}{3}\{\frac{5}{8} + \frac{5}{6} + \frac{6}{6}\} \therefore$$

$$p \cong 0.82$$

$$r = \frac{1}{3}\{\frac{5}{5} + \frac{5}{8} + \frac{6}{7}\} \therefore$$

$$r \cong 0.75$$

SEGUNDO EXEMPLO - 5 GRUPOS FORMADOS

No segundo exemplo, os 5 grupos formados possuem as configurações ilustradas pela figura 4.22.

Aplicando as fórmulas matemáticas de precisão(p) e revocação(r), tem-se:

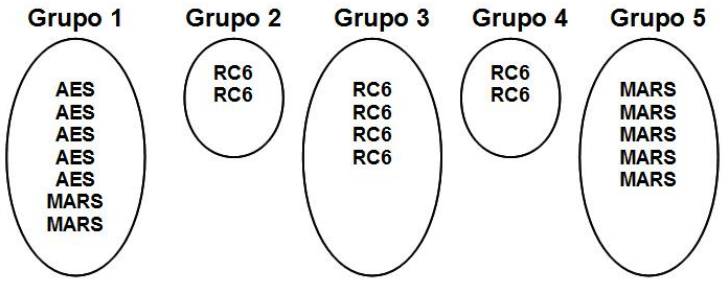


FIG. 4.22: Agrupamento com 5 grupos formados. Precisão = 0.94 e Revocação = 0.54.

$$p = \frac{1}{5}\{\frac{5}{7} + \frac{2}{2} + \frac{4}{4} + \frac{2}{2} + \frac{5}{5}\} \therefore$$

$$p \cong 0.94$$

$$r = \frac{1}{5}\{\frac{5}{5} + \frac{2}{8} + \frac{4}{8} + \frac{2}{8} + \frac{5}{7}\} \therefore$$

$$r \cong 0.54$$

TERCEIRO EXEMPLO - 5 GRUPOS FORMADOS

No terceiro exemplo, os 5 grupos formados possuem as configurações ilustradas pela figura 4.23.

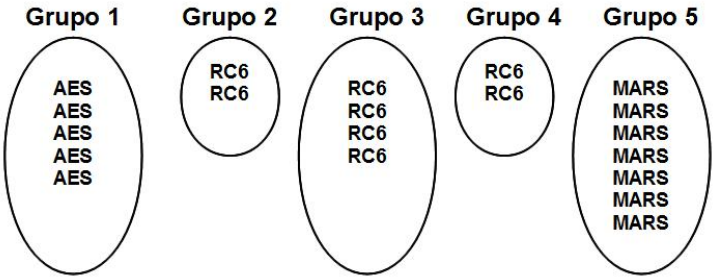


FIG. 4.23: Agrupamento com 5 grupos formados. Precisão = 1 e Revocação = 0.6.

Aplicando as fórmulas matemáticas de precisão(p) e revocação(r), tem-se:

$$p = \frac{1}{5}\{\frac{5}{5} + \frac{2}{2} + \frac{4}{4} + \frac{2}{2} + \frac{7}{7}\} \therefore$$

$$p \cong 1.0$$

$$r = \frac{1}{5}\{\frac{5}{5} + \frac{2}{8} + \frac{4}{8} + \frac{2}{8} + \frac{7}{7}\} \therefore$$

$$r \cong 0.60$$

DISCUSSÃO

Ao observar os três resultados referentes as figuras 4.21, 4.22, e 4.23, nota-se que a medida de precisão está relacionada somente com a homogeneidade [27] dos grupos. A medida de revocação está relacionada tanto com a homogeneidade dos grupo quanto com o número de grupos formados no conjunto de criptogramas.

Em uma operação de agrupamento, afirmar apenas que a precisão global foi máxima não significa necessariamente que o resultado alcançado foi satisfatório ou promissor. Por exemplo, na figura 4.24, considere uma coleção de 10 criptogramas composta por 5 textos cifrados do AES e a outra metade gerada pelo RC6. Ao se tentar agrupar esta coleção e como possível resultado aparecerem 10 grupos unitários com um criptograma cada um então, pode-se afirmar que a precisão máxima é igual a 1, em virtude de cada grupo ser formado por um uníco tipo de criptograma mas, em contrapartida, a revocação apresentará um valor baixo por causa do grande número de grupos formados. A solução correta seria a formação de dois grupos distintos com seus respectivos criptogramas, gerados pela mesma cifra. Desta forma, tanto a medida de precisão quanto de revocação são necessárias juntas para uma melhor avaliação do agrupamento formado.

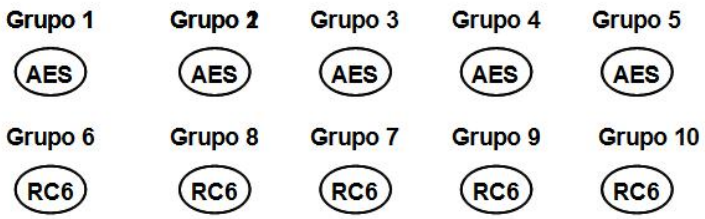


FIG. 4.24: Agrupamento com 10 grupos formados. Precisão = 1 e Revocação = 0.2.

[27] Um grupo homogêneo significa que neste grupo só existem criptogramas similares ou iguais gerados pela mesmo tipo de cifra e chave.

Aplicando as fórmulas matemáticas de precisão(p) e revocação(r), tem-se:

$$p = \frac{1}{10}\{\frac{1}{1} + \frac{1}{1} + \frac{1}{1} + \frac{1}{1} + \frac{1}{1} + \frac{1}{1} + \frac{1}{1} + \frac{1}{1} + \frac{1}{1} + \frac{1}{1}\} \therefore$$

$$p \cong 1.0$$

$$r = \frac{1}{10}\{\frac{1}{5} + \frac{1}{5} + \frac{1}{5} + \frac{1}{5} + \frac{1}{5} + \frac{1}{5} + \frac{1}{5} + \frac{1}{5} + \frac{1}{5} + \frac{1}{5}\} \therefore$$

$$r \cong 0.2$$

Convém ressaltar que o Algoritmo Genético modelado fornece nos casos testados valores máximos de precisão e revocação nas tarefas de agrupamento.

5 METODOLOGIA DE CLASSIFICAÇÃO

USO DO TEMPLATE MATCHING

O bloco binário é o parâmetro utilizado neste trabalho para uma análise de classificação. Um exemplo que demonstra a importância da análise de blocos binários em um processo de classificação de cifras pode ser visto em (NAGIREDDY, 2008). Nesse trabalho foi utilizado o método do histograma [28] cujo objetivo principal foi a captura das propriedades estatísticas das mensagens cifradas. O método do histograma mostra as variações na frequência de ocorrência dos *caracteres* dos criptogramas. Por exemplo, as figuras 5.1 e 5.2 mostram a diferença entre dois histogramas aplicados nos dois algoritmos criptográficos distintos RC5 [29] e TDES [30]

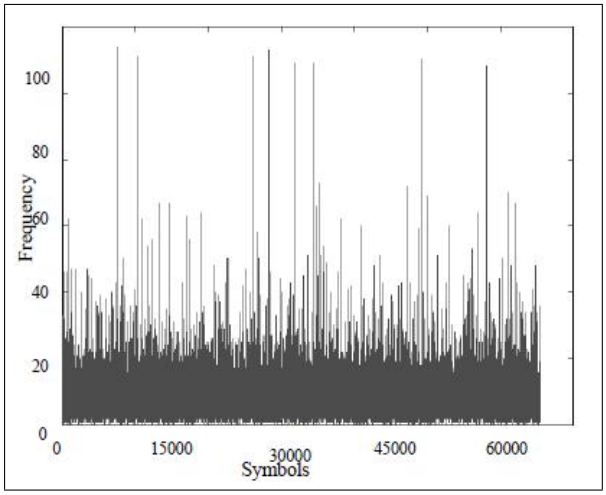


FIG. 5.1: Método do Histograma para cifra **RC5** no modo ECB (NAGIREDDY, 2008).

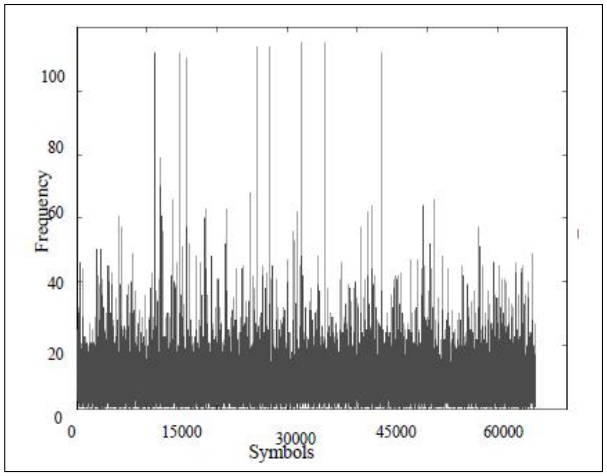


FIG. 5.2: Método do Histograma para cifra **TDES** no modo ECB (NAGIREDDY, 2008).

Assim, a diferença gráfica no histograma, acarretada pela contagem de *caracteres* dos textos cifrados, foi utilizada por (NAGIREDDY, 2008) como critério de classificação. Entretanto, o método do histograma não conseguiu detectar nenhum padrão criptográfico no AES, mesmo no modo ECB. A figura 5.3 ilustra o histograma para a cifra AES.

[28] O método histograma é uma técnica de análise de frequência de letras que foi amplamente usada para quebrar modelos de cifras clássicas no ataque por só-texto-ilegível.

[29] RC5 é uma cifra em bloco desenvolvida por (RIVEST, 1995).

[30] Criptografia Tripla do DES (Data Encryption Standard).

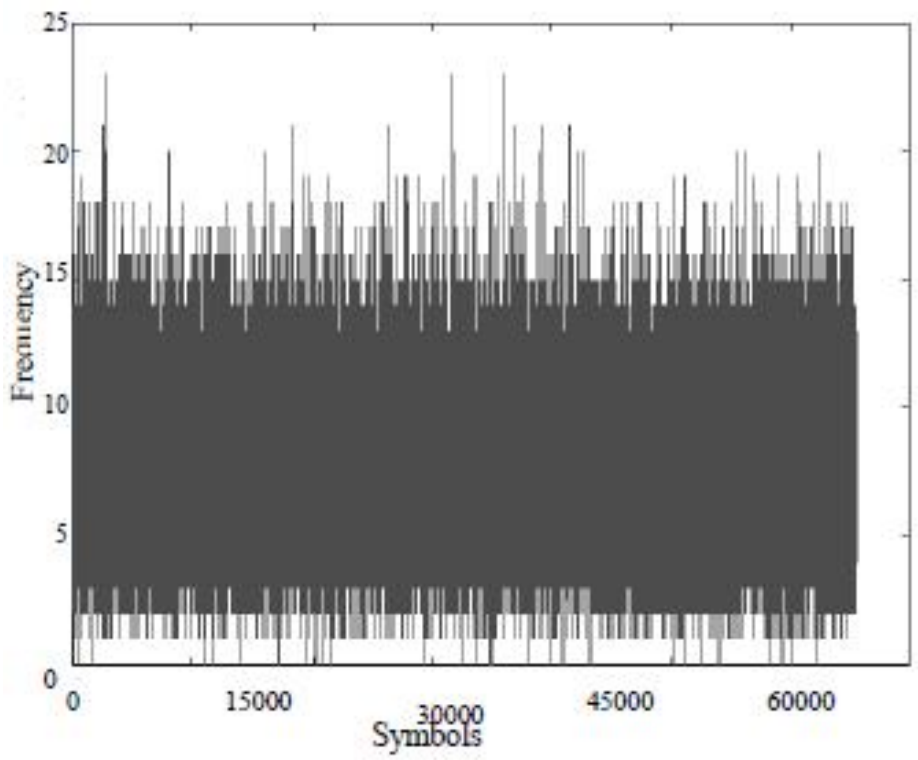


FIG. 5.3: Método do Histograma para cifra AES no modo ECB (NAGIREDDY, 2008).

A figura 5.4 ilustra a classificação de cifras utilizando o Método do Histograma. Um criptograma desconhecido gera um histograma. Este histograma é comparado com outros histogramas oriundos de criptogramas conhecidos. Deste modo, por meio de comparação gráfica (frequência de *caracteres*), um criptograma desconhecido é classificado em algum tipo de cifra conhecida.

O sistema de classificação utilizado por (NAGIREDDY, 2008) é similar ao método de classificação na área de Reconhecimento de Padrões conhecido como *Template Matching* [31].

No exemplo da figura 5.4, a regra utilizada por (NAGIREDDY, 2008) foi a comparação gráfica da frequência de *caracteres* dos diferentes tipos de histogramas armazenados com um determinado histograma proveniente de uma cifra desconhecida.

Foi associado ao Algoritmo Genético o método do *Template Matching*. O conjunto de dados previamente armazenados corresponde ao agrupamento de criptogramas conhecidos realizado pelo Algoritmo Genético. O agrupamento de criptogramas ou "dicionário" de blocos conhecidos é realizado em uma fase de treinamento. A regra arbitrada para

[31] Vide Capítulo 2.

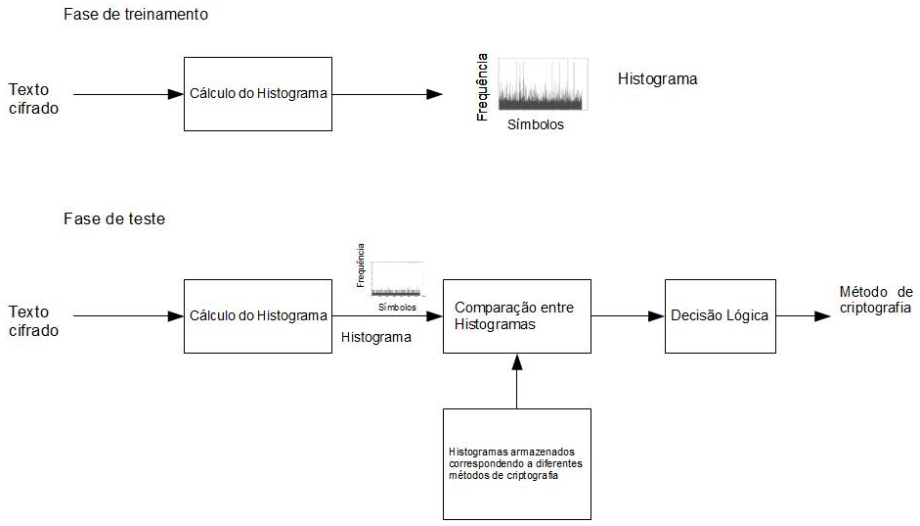


FIG. 5.4: Classificação baseada no Método do Histograma (NAGIREDDY, 2008)

a realização da classificação foi a operação de interseção de blocos binários entre o criptograma desconhecido e o "dicionário". Se houver a interseção dos blocos binários do criptograma desconhecido com algum bloco binário do agrupamento ("dicionário") então, o criptograma desconhecido será classificado no grupo onde houve a interseção de blocos. Caso contrário, se não houver interseção, o criptograma não será classificado. Se um determinado criptograma desconhecido for classificado, significa que houve a interseção de pelo menos um bloco binário com o "dicionário". Esta fase em que ocorre a possível interseção de blocos binários é conhecida como fase de teste. A figura 5.5 ilustra o sistema de classificação.

Um determinado criptograma, após ser classificado, é integrado ao conjunto de criptogramas agrupados. Os blocos binários do criptograma classificado, que não fizeram interseção com os criptogramas do conjunto de treinamento, também irão compor o "dicionário" do sistema classificador. Estes blocos binários do criptograma classificado, que não acarretaram na interseção de blocos, podem se correlacionar com outros blocos binários pertencentes a outros possíveis criptogramas similares. Isto significa "fortalecer" o conjunto de treinamento. Assim, a cada classificação, há um gradativo aumento da probabilidade de ocorrência do número de interseções de blocos binários entre criptogramas similares.

No sistema classificador do Algoritmo Genético modelado não há possibilidade de erro de classificação. O criptograma desconhecido é classificado corretamente ou não é classificado. A garantia de não haver erros de classificação decorre de dois fatos:

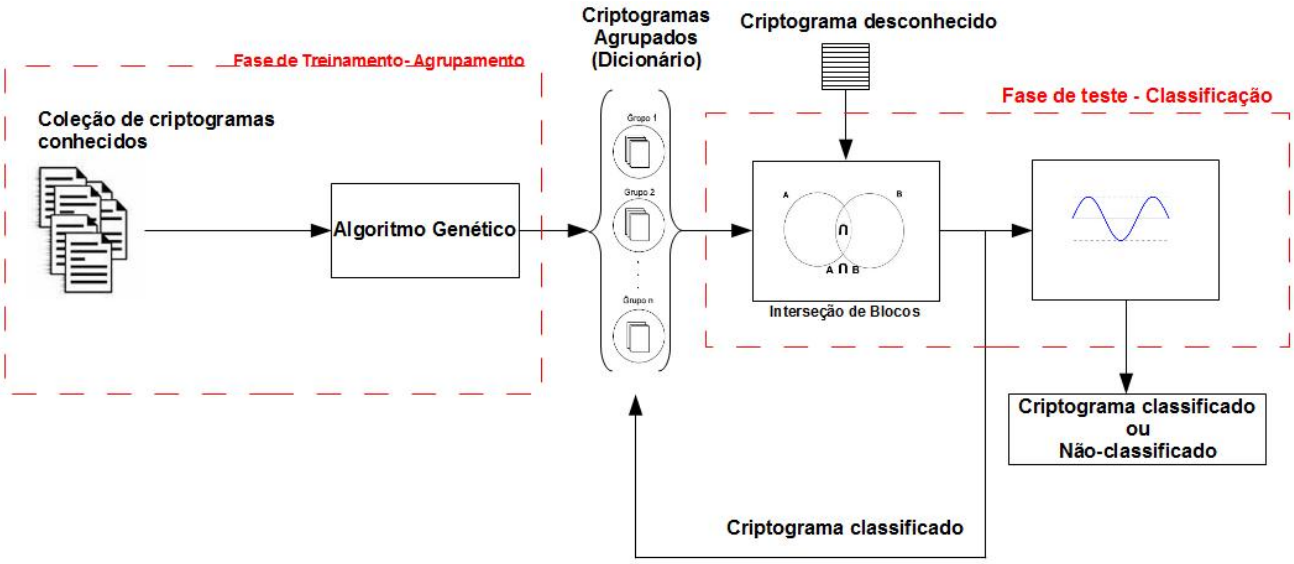


FIG. 5.5: Sistema de classificação.

a) O agrupamento realizado pelo Algoritmo Genético apresenta as métricas de revocação e precisão máximas e;

b) Criptogramas não similares, não apresentam nenhum bloco binário em comum.

A visualização gráfica da possível interseção dos blocos binários do criptograma desconhecido com o "dicionário" é feita por meio da função trigonométrica *seno*. Esta função foi adotada em virtude de sua simplicidade didática na introdução da área de Reconhecimento de Padrões em (BISHOP, 2006). Assim, sendo n_i a frequência do i-ésimo bloco binário de um criptograma então, o gráfico do "dicionário" será modelado de acordo com a função f(x) = seno$(2\pi * n_i)$. Se houver a interseção de blocos binários entre um criptograma desconhecido e o "dicionário" então, a interseção será visualizada (vide figura 5.6) por meio de uma sobreposição das curvas gráficas senoidais dos blocos binários do criptograma desconhecido e do "dicionário".

DISCUSSÃO

Na criptoanálise clássica não-computacional, os digramas [32] e trigramas eram rigorosamente analisados em cifras polialfabéticas. No trabalho desenvolvido por (NAGIREDDY,

[32]Sequência de dois carateres. Em Inglês e em outros idiomas, a frequência relativa de diversos caracteres em texto claro pode ser usada na criptoanálise de algumas cifras. Também chamado de dígrafo (STALLINGS, 2008).

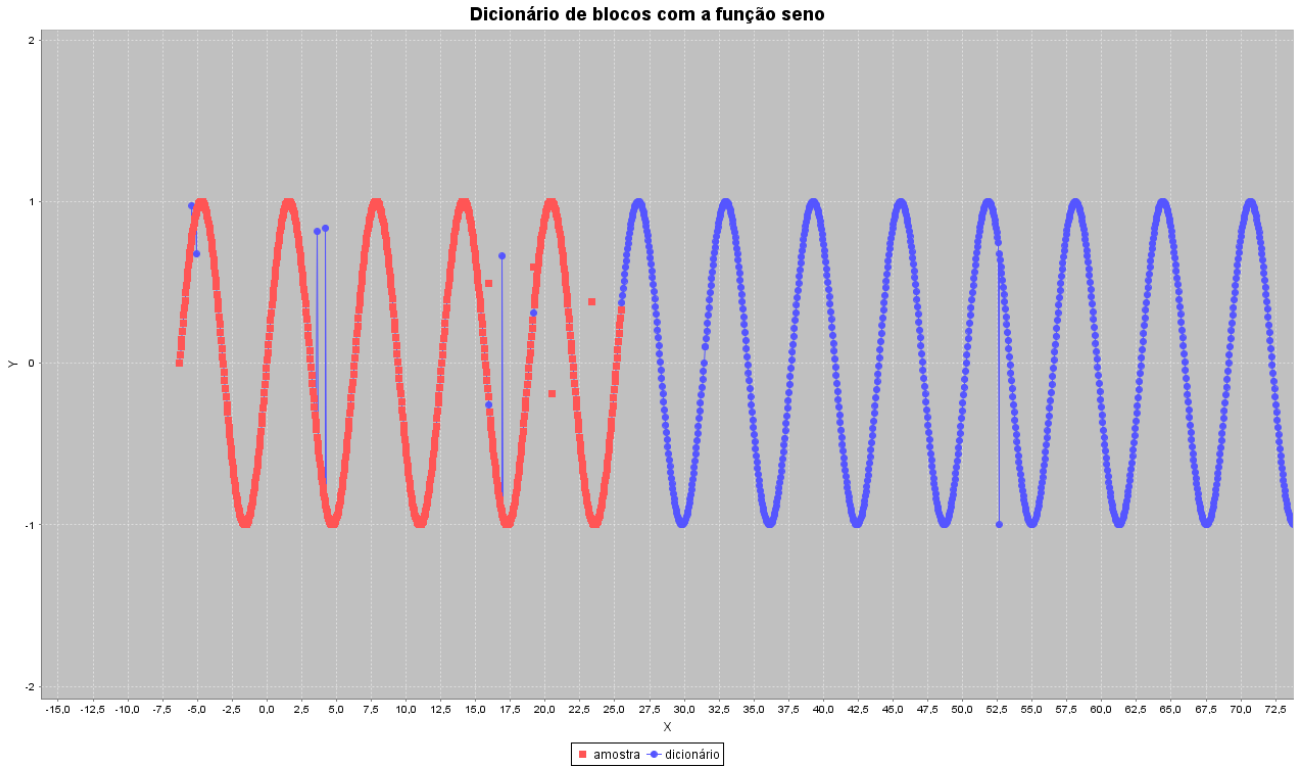


FIG. 5.6: Visualização gráfica da classificação.

2008), adotou-se a contagem unitária de *caracteres* para a aplicação do método do histograma. Cada *caractere* possui 1 byte de tamanho ou 8 bits. Deste modo, houve contagem de blocos binários de 8 bits nos criptogramas analisados. Se (NAGIREDDY, 2008) tivesse aplicado o método do histograma em blocos binários maiores do que 8 bits, talvez conseguisse detectar padrões de repetição de blocos binários no AES, no modo ECB. No sistema de classificação integrado ao Algoritmo Genético modelado, a operação de interseção é realizada com blocos binários com um tamanho de 128 bits ou 16 bytes. Relembra-se do capitulo 3, referente a fase de pré-processamento, que o tamanho de cada bloco binário pode ser determinado por qualquer divisor do tamanho da chave. Assim, foi possível detectar padrões criptográficos na cifra AES, no modo ECB, ao contrário do que aconteceu com o sistema classificador utilizado por (NAGIREDDY, 2008).

O "dicionário" do Algoritmo Génetico modelado, por exemplo, é constituído por grupos distintos de criptogramas gerados por cifras distintas. O método de correlação, entre um criptograma a priori desconhecido e o "dicionário", é a interseção de blocos binários entre eles. Havendo a interseção então, ocorre a classificação do termo desconhecido à um determinado grupo pertencente ao "dicionário". O universo de números de blocos binários

distintos disponíveis é da ordem de 2^{128} ou $3,4 \times 10^{38}$. Os blocos binários das cifras finalistas do concurso do AES não se misturaram durante os ensaios realizados nesta dissertação. Isto aparentemente pode indicar que a magnitude do universo de blocos binários de 128 bits é de fato tão grande que a probabilidade de interseção de blocos entre cifras distintas pode ser próxima de zero. Este trabalho utilizou, nos ensaios de agrupamento do capítulo 6, uma quantidade de blocos limitada a 150 criptogramas. Isto não foi suficiente para fundamentar alguma conclusão no que tange a não interseção dos blocos binários gerados por cifras distintas. Neste aspecto, talvez a execução de ensaios com uma quantidade superior a um milhão de blocos binários pudesse indicar algum conhecimento relevante.

6 EXPERIMENTOS, RESULTADOS E AVALIAÇÕES

6.1 DESCRIÇÃO DE EXPERIMENTOS

Nos experimentos foram utilizados textos claros ininteligíveis [33] e textos claros procedentes da Bíblia inglesa (*www.o-bible.com*) com os tamanhos de 10, 8 e 6 Kbytes. O tamanho da chave utilizada foi de 128 bits para todos os ensaios. Os algoritmos criptográficos utilizados foram os cinco finalistas do concurso do AES promovido pelo NIST.

6.1.1 PRIMEIRO CONJUNTO DE EXPERIMENTOS

6.1.1.1 ENSAIO COM ALGORITMOS CRIPTOGRÁFICOS DISTINTOS COM A MESMA CHAVE

Este ensaio tem como propósito realizar o agrupamento automático de uma coleção de criptogramas gerados por cinco algoritmos criptográficos diferentes (AES, MARS, SERPENT, TWOFISH e RC6). Foram usados 30 textos claros com 10 Kbytes de tamanho, extraídos da Bíblia inglesa, totalizando 150 criptogramas. A figura 6.1 abaixo ilustra graficamente o resultado do agrupamento automatizado.

[33] Os textos claros ininteligíveis são compostos por vocábulos do idioma *latim* distribuidos de forma pseudo-aleatória. Os textos foram gerados no site *http://pt.lipsum.com*.

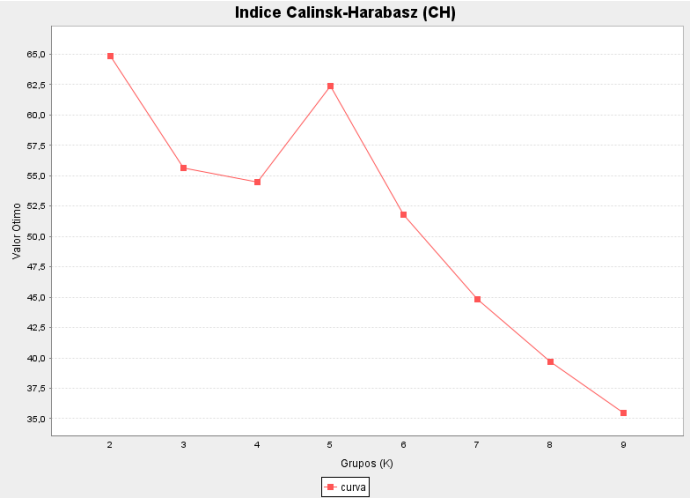


FIG. 6.1: Agrupamento automático com 5 cifras distintas e uma chave comum. Para k igual a 5 temos o número correto de grupos.

6.1.1.2 ENSAIO COM ALGORITMO CRIPTOGRÁFICO AES COM CHAVES DISTINTAS

Este ensaio tem o objetivo de realizar o agrupamento automático de um conjunto de criptogramas gerados somente pelo AES. Este algoritmo utilizou cinco chaves distintas e o mesmo conjunto de textos claros do ensaio do subitem 6.1.1.1. A figura 6.2 ilustra graficamente o resultado do agrupamento.

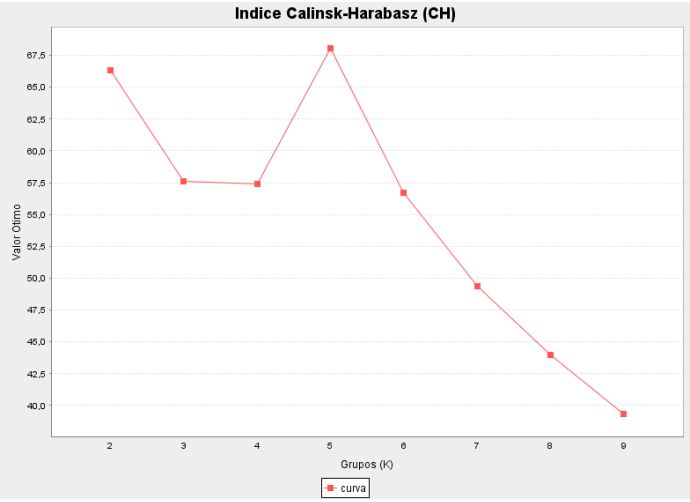


FIG. 6.2: Agrupamento automático de criptogramas gerados somente pelo AES, utilizando 5 chaves distintas. Para k igual a 5 temos o número correto de grupos. Neste caso, o "joelho" que corresponde ao ponto máximo global da função de avaliação indica o correto particionamento.

6.1.1.3 RESULTADOS E AVALIAÇÕES

Em relação ao ensaio do subitem 6.1.1.1, o "joelho" da curva gráfica da figura 6.1 confirma o número correto de grupos da coleção de criptogramas porque este resultado mostra a existência de "assinatura" própria para cada tipo de algoritmo criptográfico utilizado para cifrar uma mesma coleção de textos claros.

Analisando a figura 6.2 do ensaio do subitem 6.1.1.2, o resultado do número correto de grupos aparece também no "joelho" do gráfico que, neste caso, também é o "ponto máximo" da curva gráfica. Assim, é verificado que cada chave distinta impõe uma característica própria de "assinatura".

Estas duas experiências foram importantes para demonstrar que tanto chaves como as cifras distintas apresentaram um padrão criptográfico.

6.1.2 SEGUNDO CONJUNTO DE EXPERIMENTOS

6.1.2.1 ENSAIO COM ALGORITMOS CRIPTOGRÁFICOS DISTINTOS COM A MESMA CHAVE - INFLUÊNCIA DO TAMANHO DO TEXTO CLARO

Este experimento é similiar ao ensaio do subitem 6.1.1.1 e tem como objetivo analisar o resultado do agrupamento automático em um conjunto de criptogramas com tamanhos menores. Foram utilizados textos cifrados de 8 Kbytes e 6 Kbytes para a análise. As figuras 6.3 e 6.4 mostram graficamente a saída do índice CH.

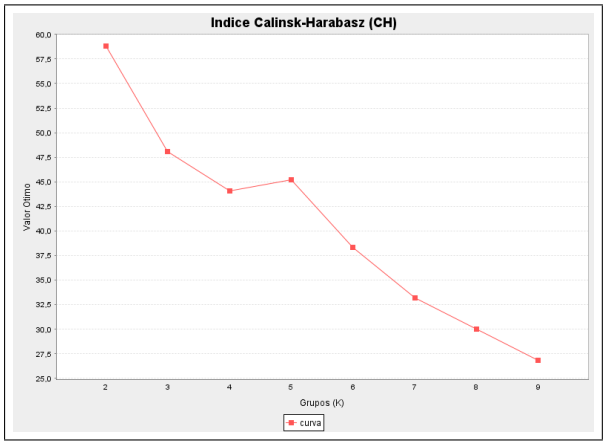


FIG. 6.3: Agrupamento automático com textos cifrados de **8 Kbytes**. Cinco algoritmos criptográficos distintos e uma chave comum em uso.

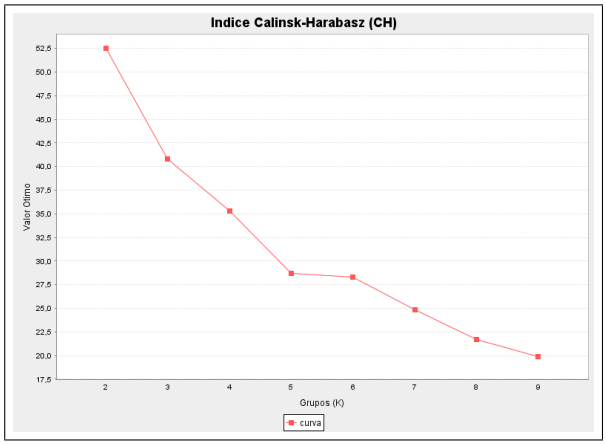


FIG. 6.4: Agrupamento automático com textos cifrados de **6 Kbytes**. Cinco algoritmos criptográficos distintos e uma chave comum em uso.

6.1.2.2 ENSAIO COM ALGORITMO CRIPTOGRÁFICO AES COM CHAVES DISTINTAS - INFLUÊNCIA DO TAMANHO DO TEXTO CLARO

Este experimento é similiar ao ensaio do subitem 6.1.1.2 e tem como objetivo analisar o resultado do agrupamento automático em um conjunto de criptogramas com tamanhos menores. Foram utilizados textos cifrados de 8 e 6 Kbytes. As figuras 6.5 e 6.6 mostram graficamente a saída do índice CH.

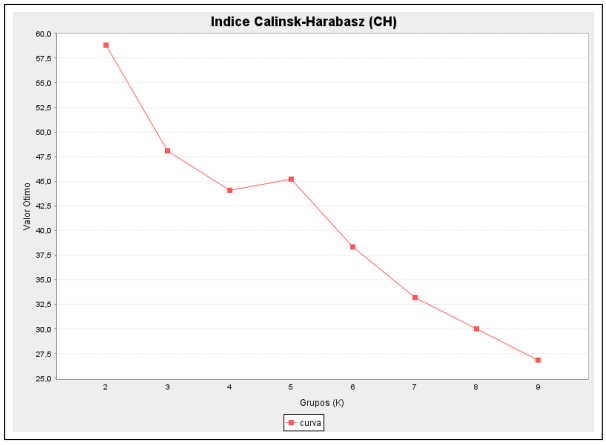


FIG. 6.5: Agrupamento automático com textos cifrados de **8 Kbytes** de tamanho. AES com 5 chaves distintas.

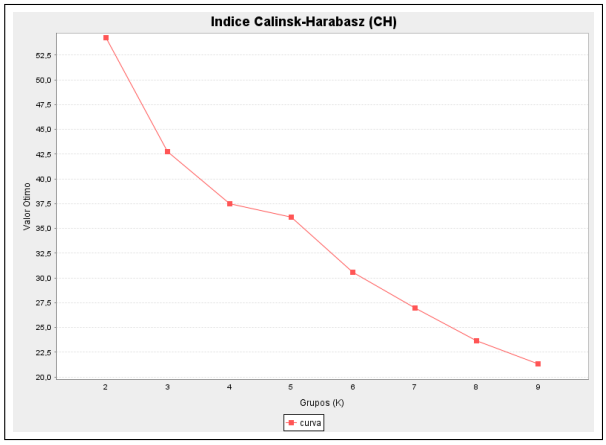


FIG. 6.6: Agrupamento automático com textos cifrados de **6 Kbytes** de tamanho. AES com 5 chaves distintas.

6.1.2.3 RESULTADOS E AVALIAÇÕES

Com base na figura 6.3, o resultado encontrado para textos de 8 Kbytes foi similar ao gráfico da figura 6.1. Com a redução do tamanho do criptograma para 8 Kbytes, há também a redução do número de blocos binários a serem analisados. Assim, pode-se perceber que o valor ótimo do índice CH da figura 6.3 reduziu em comparação ao gráfico da figura 6.1. Os criptogramas similares com tamanhos menores possuem menos blocos binários comuns entre si. Deste modo, a similaridade entre os textos cifrados diminui fazendo com que o valor da qualidade do melhor *cromossomo* do Algoritmo Genético modelado também diminua.

A figura 6.4 mostra o resultado para textos cifrados de 6 Kbytes. O resultado foi considerado insatisfatório devido ao fato do gráfico não indicar o número correto de grupos. Com a redução do tamanho do criptograma para 6 Kbytes, pode-se notar que o índice CH não gerou nenhum " joelho" na curva gráfica.

Os resultados dos ensaios envolvendo algoritmos distintos estabeleceu um limite inferior de 8 Kbytes para cada criptograma. Convém ressaltar que este limite está correlacionado ao agrupamento automatizado, proporcionado pela função *Calinski-Harabasz* (CH).

Nos ensaios realizados somente com o AES, figuras 6.5 e 6.6, estabeleceu-se um limite de 8 Kbytes para o tamanho dos criptogramas. Isto devido ao fato da curva gráfica da figura 6.6 não fornecer nenhuma indicação sobre o número correto de grupos.

6.1.3 TERCEIRO CONJUNTO DE EXPERIMENTOS

6.1.3.1 ENSAIO COM ALGORITMOS CRIPTOGRÁFICOS DISTINTOS COM A MESMA CHAVE - UTILIZANDO TEXTOS CLAROS ININTELIGÍVEIS DO IDIOMA LATIM

Este experimento tem por objetivo verificar se os textos claros ininteligíveis do idioma latim tem alguma influência na "assinatura" imposta pelos diferentes tipos de algoritmos criptográficos. A figura 6.7 mostra o resultado para textos cifrados de 10 Kbytes.

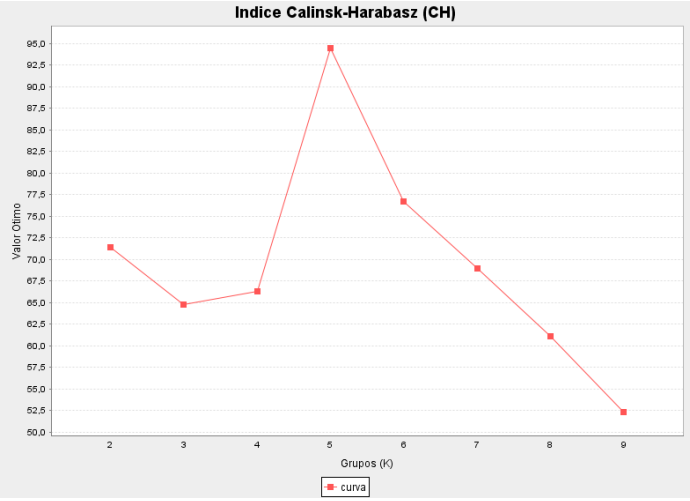


FIG. 6.7: Agrupamento automático com textos cifrados de **10 Kbytes**. Cinco algoritmos criptográficos distintos e uma chave comum. Número correto de grupos para k igual a 5.

6.1.3.2 RESULTADOS E AVALIAÇÕES

A utilização de textos claros pseudo-aleatórios com 10 Kbytes de tamanho no ensaio do item 6.1.3.1, mostrou que não houve nenhuma influência da mudança do tipo de texto claro nos resultados gráficos em relação ao experimentos anteriores.

6.1.4 QUARTO CONJUNTO DE EXPERIMENTOS

6.1.4.1 ENSAIO COM ALGORITMOS CRIPTOGRÁFICOS DISTINTOS COM A MESMA CHAVE - INFLUÊNCIA DO TAMANHO DO TEXTO CLARO ININTELIGÍVEL NO IDIOMA LATIM

Este experimento tem o propósito verificar a influência da redução do tamanho dos arquivos cifrados no gráfico gerado pelo índice CH. As figuras 6.8 e 6.9 mostram os resultados gráficos encontrados para textos cifrados de 8 Kbytes e 6 Kbytes, respectivamente.

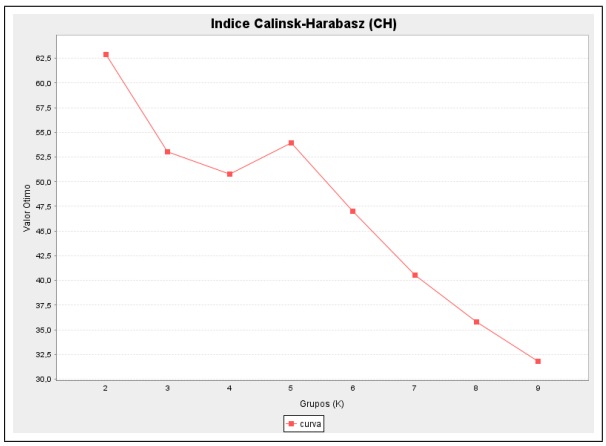


FIG. 6.8: Agrupamento automático com textos cifrados de **8 Kbytes**. Algoritmos distintos e uma chave comum.

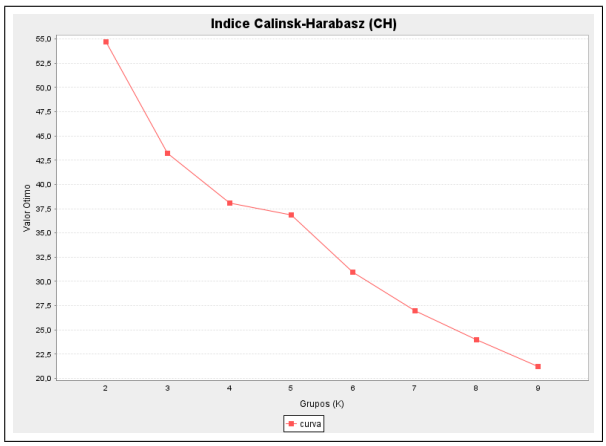


FIG. 6.9: Agrupamento automático com textos cifrados de **6 Kbytes**. Algoritmos distintos e uma chave comum.

6.1.4.2 RESULTADOS E AVALIAÇÕES

Para ensaios realizados com algoritmos distintos, os resultados apresentados nas figuras 6.8 e 6.9, indicam um limite inferior de 8 Kbytes de tamanho para os textos claros ininteligíveis do idioma latim.

6.1.4.3 DISCUSSÃO

Os resultados observados para o agrupamento automático de criptogramas gerados por cifras distintas utilizando textos claros ininteligíveis do idioma latim foram similares aos ensaios que utilizaram textos claros da língua inglesa. Isto mostrou que a "assinatura" imposta por cada tipo diferente de algoritmo criptográfico não dependeu do tipo de texto claro.

6.1.5 ENSAIO DE CLASSIFICAÇÃO

Para realização deste ensaio, as amostras da seção 6.1.1.1 foram utilizadas como conjunto de treinamento para a formação do "dicionário" binário (vide capítulo 5). Um criptograma, a priori desconhecido é comparado com este "dicionário" por meio de uma regra arbitrada que neste caso é a interseção de blocos binários entre os criptogramas do conjunto de treinamento e o próprio criptograma desconhecido. Havendo a interseção então, o criptograma desconhecido é classificado. A visualização da interseção de blocos é mostrada por meio de uma sobreposição gráfica. O gráfico senoidal azul representa os blocos binários da amostra desconhecida enquanto que o gráfico vermelho corresponde aos blocos binários do "dicionário". O criptograma escolhido para representar a amostra desconhecida é uma amostra gerada pela cifra Twofish e que também utilizou a mesma chave criptográfica do "dicionário". A figura 6.10 ilustra a operação de interseção.

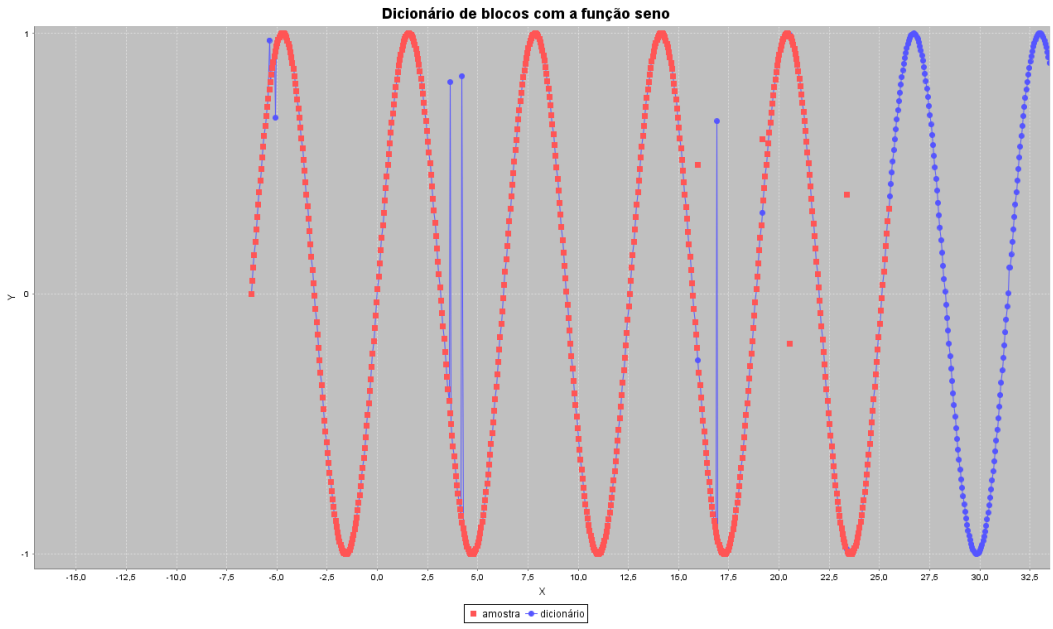


FIG. 6.10: Gráfico de visualização de classificação.

DISCUSSÃO

A implementação da metodologia do *Template Matching* no Algoritmo Genético possui uma complexidade baixa em virtude de utilizar arbitrariamente apenas a regra de interseção de blocos binários. Conforme visto no capitulo 5, a análise e contagem de blocos binários é importante na tarefa de classificação de criptogramas. Para aumentar ainda mais a capacidade de classificação do Algoritmo Genético modelado, faz-se mister utilizar um "dicionário" formado com muitos tipos de cifras e chaves. (DILEEP, 2006) utilizou "dicionários" específicos para a tarefa de classificação e, por conseguinte, conseguiu resultados melhores do que tivesse usado um único "dicionário" comum.

7 COMPARAÇÃO DA TÉCNICA COM OUTRAS FERRAMENTAS MODELADAS

Este capítulo tem como objetivo mostrar as diferenças entre os resultados obtidos pelo Algoritmo Genético modelado em comparação as técnicas de Agrupamento Hierárquico aplicadas por (CARVALHO, 2006) e (SOUZA, 2007) e a técnica de classificação, por meio do histograma, utilizada por (NAGIREDDY, 2008).

7.1 TÉCNICAS DE AGRUPAMENTO HIERÁRQUICO E HISTOGRAMA *VERSUS* ALGORITMO GENÉTICO MODELADO

7.1.1 AGRUPAMENTO HIERÁRQUICO

Nos ensaios realizados por (CARVALHO, 2006) e (SOUZA, 2007) na aplicação das técnicas de Agrupamento Hierárquico em uma coleção de criptogramas, havia a necessidade do conhecimento da quantidade de grupos realmente existentes. Além dessa informação, a partição dos grupos no processo de agrupamento tinha que ser feita por meio de um parâmetro de corte que representava o nível de similaridade. Por exemplo, a figura 7.1 ilustra um gráfico tipo dendrograma [34] que fornece um resultado possível de agrupamento de criptogramas. Existe a formação de três grupos, em que a similaridade entre os criptogramas de grupos diferentes é sempre zero. Caso fosse desejado apenas dois grupos dos três realmente existentes então, poderá haver a formação de dois grupos com criptogramas gerados por chaves distintas.

No Algoritmo Genético modelado não existe a necessidade da informação prévia da quantidade correta de grupos existentes e nem a inserção de algum parâmetro de similaridade para a realização do agrupamento. Por meio função de avaliação *Calisnki-Harabazs (CH)* [35] foi possivel automatizar o agrupamento de criptogramas gerados por chaves ou cifras distintas.

[34] Estrutura gráfica que representa os agrupamentos realizados, respeitando a ordem de união dos objetos e o nível de similaridade em que cada uma ocorreu (JAIN, 1999).

[35] Vide capítulo 4.

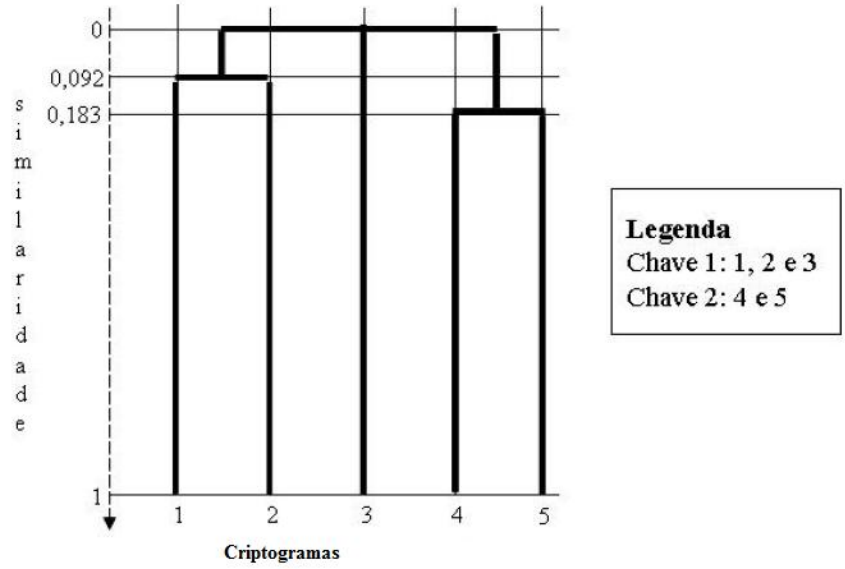


FIG. 7.1: Possível resultado de agrupamento (CARVALHO, 2006).

7.1.2 CLASSIFICAÇÃO - TÉCNICA DO HISTOGRAMA

(CARVALHO, 2006) aplicou a técnica de agrupamentos hieráquicos somente para o particionamento de criptogramas e, posteriormente, (SOUZA, 2007) dando continuidade aos trabalhos de agrupamento, tentou implementar um processo de classificação por meio da utilização das redes neuronais [36]. Os resultados mostraram-se insatisfatórios pelo fato de que em alguns ensaios, criptogramas cifrados com chaves diferentes foram agrupados no mesmo grupo. Neste contexto, nos trabalhos desenvolvidos tanto por (CARVALHO, 2006) e (SOUZA, 2007), não houve sucesso na tarefa de classificação.

(NAGIREDDY, 2008) utilizou a técnica do histograma da área de Reconhecimento de Padrões para a classificação de cifras de blocos [37]. Nesse trabalho, foi afirmado que não era possível encontrar padrões criptográficos no AES, no modo ECB. O Algoritmo Genético modelado [38] realizou ensaios satisfatórios de agrupamento de criptogramas gerado com a cifra AES variando o tipo de chave (vide capítulo 6). Desta forma, há padrões detectados no AES que permitem a classificação de acordo com a metodologia adotada nesta dissertação.

[36] Foi utilizado o mapa de Kohonen.

[37] Vide capítulo 5.

[38] Cabe ressaltar que a expressão "Algoritmo Genético modelado" neste trabalho faz referência ao Algoritmo Genético agrupador integrado a técnica de classificação "Template Matching".

7.2 DISCUSSÃO

Em todos os resultados dos agrupamentos de criptogramas executados pelo Algoritmo Genético modelado, para o Banco de Dados analisado, observou-se valores de revocação e precisão máximas a partir de textos com 8 Kbytes. Do capitulo 4, relembra-se que os valores de precisão estão relacionados somente a homogeneidade do grupo formado e revocação aos números de grupos formados e homogeneidade. Este tipo de resultado nem sempre acontecera nos ensaios de (CARVALHO, 2006) e (SOUZA, 2007).

Outro fato observado é que Algoritmo Genético modelado utilizou 150 criptogramas em todos os seus ensaios enquanto que (CARVALHO, 2006) e (SOUZA, 2007) utilizaram 1500 criptogramas para mostrar resultados satisfatórios. A figura 7.2 ilustra a quantidade de amostras utilizadas por (CARVALHO, 2006) e (SOUZA, 2007) em um ensaio aplicado no AES.

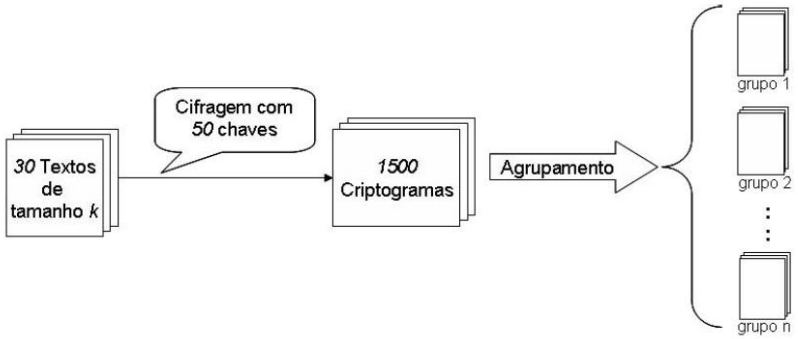


FIG. 7.2: Quantidade de amostras utilizadas (CARVALHO, 2006) (SOUZA, 2007).

O Algoritmo Genético modelado utilizou 10% da quantidade de chaves e textos usados por (CARVALHO, 2006) e (SOUZA, 2007) para conseguir o agrupamento satisfatório de criptogramas gerados pelo AES. A figura 7.3 ilustra a quantidade de amostras utilizadas por este trabalho. Outro fato também importante é que esta dissertação analisou as cifras MARS, RC6, AES, Serpent e Twofish enquanto que os trabalhos supramencionados, em relação aos cinco finalistas do concurso do AES, se limitaram somente ao estudo do AES [39].

(CARVALHO, 2006) e (SOUZA, 2007) desenvolveram um trabalho inovador no IME sobre Reconhecimeto de Padrões criptográficos utilizando as técnicas de Agrupamento Hierárquico. Estas técnicas fazem parte de uma taxonomia de agrupamento preconizada

[39] O Algoritmo criptográfico Rinjdael, vencedor do concurso do NIST.

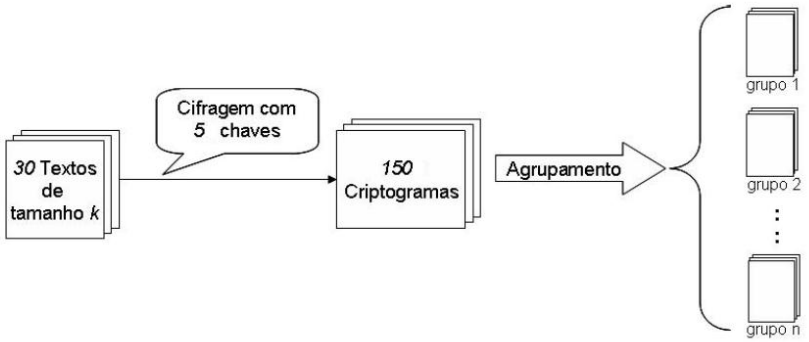


FIG. 7.3: Quantidade de amostras utilizadas pelo Agoritmo Genético modelado.

por (JAIN, 1999), de acordo com a figura 7.4.

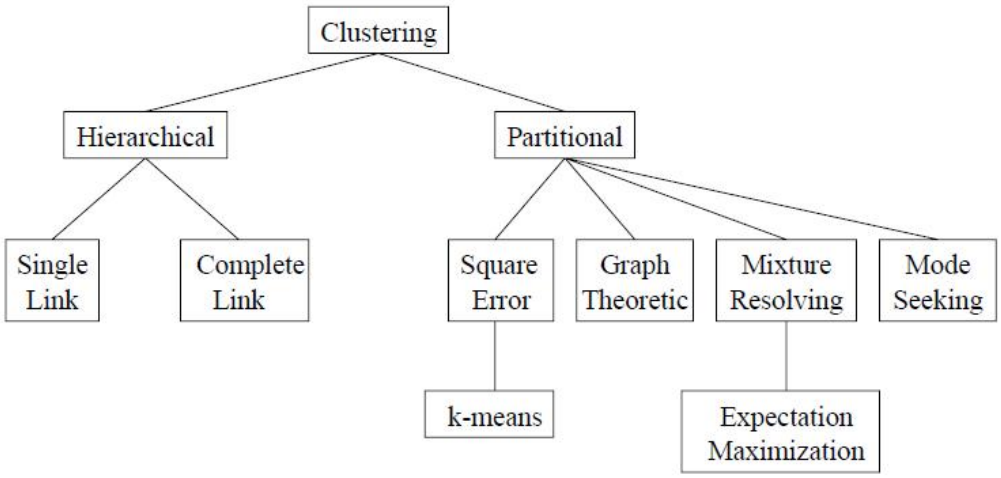


FIG. 7.4: Taxonomia das abordagens de agrupamento (JAIN, 1999).

(TORRES, 2010) no IME, utilizando o outro ramo da taxonomia ilustrada na figura 7.4, aplicou a Teoria dos Grafos para a realização de agrupamentos de criptogramas gerados por chaves ou cifras distintas. Os grafos pertecem as técnicas de Agrupamento Não-hierárquico ou Particionais. Como exemplo, a figura 7.5 ilustra o resultado de um agrupamento realizado pela Teoria dos Grafos em um conjunto de criptogramas. Foram agrupados 105 criptogramas em 5 grupos distintos. Cada grupo é formado por criptogramas gerados por cifras [40] distintas. Pode-se observar nos grafos da figura 7.5 que eventualmente ocorrerá que a similaridade entre dois criptogramas seja igual a zero, embora tenham sido cifrados com a mesma chave. Mesmo assim, esses criptogramas estão no

[40]Foram utilizadas as 5 cifras finalistas do concurso do AES.

mesmo grupo. Neste caso, a pertinência ao grupo é fornecida pela co-similaridade com um terceiro criptograma.

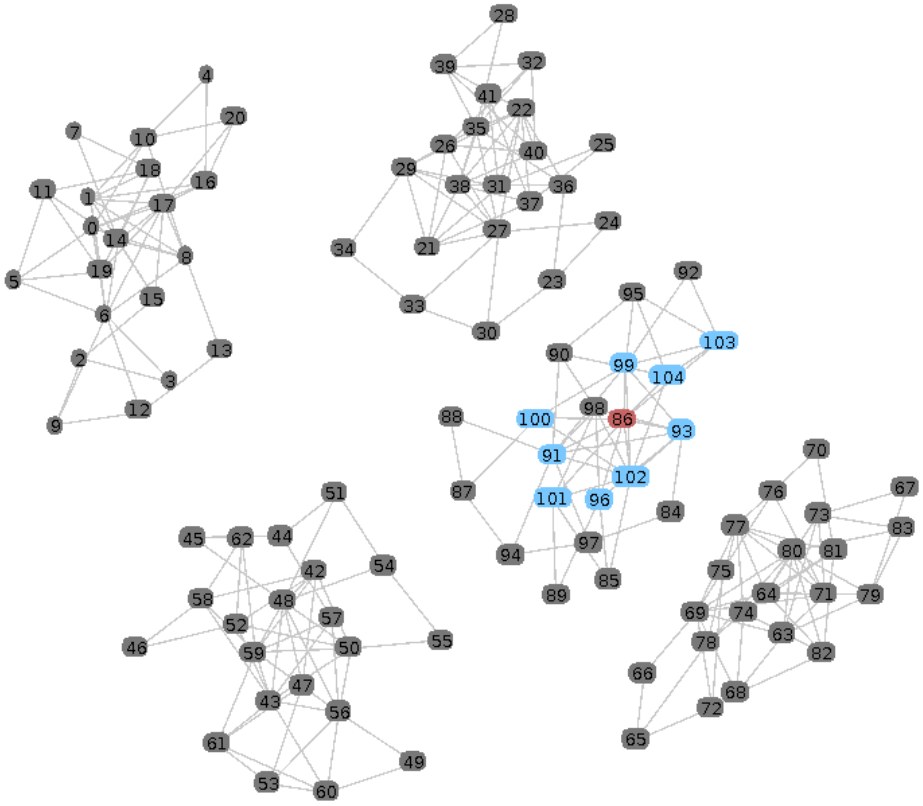


FIG. 7.5: Agrupamento realizado por Grafos (TORRES, 2010).

O desempenho nos resultados de agrupamento realizados pelos Grafos foram similares aos alcançados pelo Algoritmo Genético modelado. Entretanto, cabe ressaltar que os Grafos realizam apenas o agrupamento. Trabalhos como (BANDYOPADHYAY, 2002),(BANDYOPADHYAY, 2007) entre outros, realizaram tarefas de agrupamento aplicando o Algoritmo Genético como ferramenta todavia, utilizaram banco de dados não correlacionados a criptogramas.

8 CONCLUSÕES

8.1 CONSIDERAÇÕES FINAIS

Os testes estatísticos propostos pelo NIST têm por objetivo a detecção de "assinaturas" nos criptogramas gerados pelos algoritmos ensaiados e, assim, testar se os mesmos são bons geradores de números pseudo-aleatórios. Entretanto, conforme visto nesta dissertação, no caso dos algoritmos finalistas do processo para a escolha do AES, esses testes não foram suficientes para detectar os padrões que vários autores, posteriormente e com diferentes técnicas, foram capazes de detectar. Desta forma, os resultados mostrados pelo Algoritmo Genético modelado, questionam a validade dos testes estatísticos do NIST como requisitos que estabelecem níveis de segurança de algoritmos criptográficos. É necessário, portanto, que a metodologia atual de certificação seja acrescida de novos ensaios para aumentar o nível de confiabilidade de seus resultados. As técnicas de Reconhecimento de Padrões são uma alternativa que demonstra utilidade. A técnica de agrupar e classificar do Algoritmo Genético modelado assim como as outras técnicas desenvolvidas no IME por (CARVALHO, 2006) e (SOUZA, 2007), podem assessorar as autoridades militares na aquisição comercial ou elaboração de um algoritmo criptográfico suficientemente seguro no âmbito militar.

Convém ressaltar que o Algoritmo Genético é utilizado juntamente com a técnica do Template Matching, o qual foi capaz de separar e classificar corretamente os criptogramas gerados pelos cinco algoritmos finalistas do concurso do AES: MARS, RC6, Rijndael, Serpent e Twofish; o que leva a classificação desses algoritmos a partir dos criptogramas gerados pelos mesmos. A classificação relatada demonstra a existência de padrões 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. A técnica apresentada contribui principalmente na classificação correta de cifras, apresentado melhores resultados na identificação de cifras do que os trabalhos anteriores relacionados relatados no IME (CARVALHO, 2006) e (SOUZA, 2007) e pelo Instituto Indiano de Tecnologia de Madras (NAGIREDDY, 2008). 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 criptográficos devem ter força suficiente para resistir a ataques sob a condição de "pior caso".

Assim como o Algoritmo Genético modelado pode ser utilizado para o assessoramento da aquisição comercial de alguma cifra de bloco também pode ser aplicado em uma fase preliminar de um ataque por "Só-texto-ilegível" com o objetivo de reduzir o esforço empregado por um criptoanalista [41]

Dois fatos importantes a destacar no capitulo 5 que trata da metodologia de classificação foram: O comportamento dos resultados obtidos no agrupamento de criptogramas gerados por cifras distintas. O espaço binário, no universo de grandeza 2^{128} bits, é aparentemente tão grande que os blocos binários, gerados pelas cifras finalistas do concurso do AES, parecem não se encontrarem em uma mesma cifra. O outro fato foi a afirmação equivocada de (NAGIREDDY, 2008): "Criptogramas gerados pelo AES, no modo ECB, são comparativamente mais seguros [42]". A técnica do Algoritmo Genético desta dissertação e os trabalhos de (CARVALHO, 2006) e (SOUZA, 2007), provaram o contrário.

8.2 CONTRIBUIÇÕES DO TRABALHO

As maiores contribuições desta tese foram:

- Melhoramento no desempenho do agrupamento de criptogramas;
- A implementação de um método de classificação de criptogramas;
- Desenvolvimento de uma ferramenta de agrupamento sem a necessidade de se conhecer o número exato de grupos a serem formados; e
- Desenvolvimento de um trabalho focado na detecção de padrões nas 5 cifras finalistas do concurso do AES.

[41] Vide Capitulo 2 - Tipos de ataques criptoanaliticos.

[42] O AES foi comparado com as cifras DES, TDES, RC5 e Blowfish. Todas estas cifras, exceto o AES, apresentaram padrões detectados no método do histograma.

8.3 TRABALHOS FUTUROS

Os resultados desta dissertação sugerem como trabalhos futuros:

- A identificação e separação de classes de chaves para um mesmo algoritmo criptográfico;
- Estudos para modificar as transformações matemáticas nos algoritmos testados neste artigo para que, mesmo no modo ECB, estes não propaguem informações dos textos claros para os criptogramas gerados;
- Estudos na variação dos diferentes valores que foram arbitrados nos parâmetros (tamanho da população, taxas de crossover e mutação) do Algoritmo Genético;
- Estudos para a realização de agrupamento e classificação utilizando uma quantidade superior de criptogramas na ordem de grandeza de 10^6 com o objetivo de verificar uma possível ou não interseção de blocos binários entre criptogramas gerados por cifras distintas.

9 REFERÊNCIAS BIBLIOGRÁFICAS

BANDYOPADHYAY, S. e MAULIK, U. **Performance Evaluation of Some Clustering Algorithms and Validity Indices.** *IEEE Transactions on Pattern Analysis and machine Inteligence, Vol. 24, No. 12, December 2002*, 2002.

BANDYOPADHYAY, S. e SANKAR, K. ***Classification and Learning Using Genetic Algorithms.*** Springer., 2007.

BASSAB, W. e MIAZAKI, S. **Introdução à Análise de Agrupamentos.** *Associação Brasileira de Estatistica, ABE. Simpósio Nacional de Probabilidade e Estatistica, São Paulo.*, 1990.

BENITS, W. **Sistemas criptograficos baseados em identidades pessoais.** Dissertação de Mestrado, Instituto de Matematica e Estatistica da Universidade de Sao Paulo., 2003.

BISHOP, C. ***Pattern Recognition. Machine Learning***. Springer., 2006.

CALINSKI, T. e HARABASZ, J. **A dendrite method for cluster analysis.** 1974.

CARVALHO, C. A. B. **O uso de técnicas de recuperação de informações em criptoanálise**. Dissertação de Mestrado, Instituto Militar de Engenharia, 2006.

CARVALHO, L. **Datamining - A mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração.** *Ciência Moderna, Terceira edição*, 2005.

CLARK, A., . D. E. **A Parallel Genetic Algorithm for Cryptanalysis of the Polyalphabetic Substitution Cipher.** . *Cryptologia, 21 (2), 129-138*, 1997.

CLARK, A. **Modern optimisation algorithms for cryptanalysis.** Em *In Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems*, págs. 258–262, 1994.

CONCI, A. E OLIVEIRA, N. E. **Clusterização Automática na Redução da Dimensionalidade dos Dados.** *Simpósio de Pesquisa Operacional e Logistica da Marinha.*, 2008.

DELMAN, B. **Genetic Algorithms in Cryptography.** Dissertação de Mestrado, Institute of Technology Rochester, New York., 2004.

DENNING, D. ***Cryptography and data security***. *Company, USA*. Addison-Wesley Publishing, 1982.

DIFFIE, W., H. M. **New Directions in Cryptography**. *IEEE International Symposium on Information Theory, Sweden*, 1976.

DILEEP, A. D. e SEKHAR, C. C. **Identification of Block Ciphers using Support Vector Machines**. *International Joint Conference on Neural Networks Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada July 16-21*, 2006.

FREI, F. ***Introdução à análise de agrupamentos. Teoria e prática.*** Unesp., 2006.

GARG, P. **Genetic Algorithm Attack on Simplified Data Encryption Standard Algorithm.** *International journal Research in Computing Science*, 2006.

GOLDSCHMIDT, R. E PASSOS, E. ***Data mining: um guia prático. Rio de Janeiro.*** Elsevier., 2005.

GRUNDLINGH, W. & VAN VUUREN, J. H. **Using Genetic Algorithms to Break a Simple Cryptographic Cipher.** 2002.

HALKIDI, M., B. Y. V. M. **Clustering algorithms and validity measures.** 2001.

HARTIGAN, J. ***Clustering Algorithms Series in Probability and Mathematical Statistics***. John Wiley & Sons., 1975.

JAIN, A. K; MURTY, M. N. F. P. J. **Data Clustering: A Review.** acm computing surveys, vol. 31, no 3, september 1999. p. 264-323. 1999.

JANSSEN, M., S. R. N. B. . K. M. **Use of a genetic algorithm in the cryptanalysis of simple substitution ciphers.** *Cryptologia, 17(1), 31- 44.*, 1993.

KNUDSEN, L. R. e MEIER., W. **Correlations in RC6 with a Reduced Number of Rounds**. *Department of Informatics, University of Bergen, N-5020 Bergen 2 FH-Aargau, CH-5210 Windisch*, 2000.

LAMBERT, J. D. A. **Cifrador simétrico de blocos: projeto e avaliação.** 2004. 353 f. dissertação (mestrado em sistemas e computação). Dissertação de Mestrado, Instituto Militar de Engenharia, Rio de Janeiro., 2004.

LI, X., R. R. M. S. e GRAVES, S. **Storm Clustering for Data-driven Weather Forecasting.** University of Alabama in Huntsville. 2008.

LIN, FENG-TSE, . K. C.-Y. **A genetic algorithm for ciphertext-only attack in cryptanalysis.** Em *In IEEE International Conference on Systems, Man and Cybernetics*, págs. 650–654, vol. 1, 1995.

LINDEN, R. ***Algoritmos Genéticos. Uma importante ferramenta da inteligência computacional.*** Segunda edição, Brasport, 2008.

MARQUES, J. S. ***Reconhecimento de Padrões. Segunda edição***. IST- Press, 2005.

MATTHEWS, R. **The use of genetic algorithms in cryptanalysis.** *Cryptologia, 17(4), 187-201.*, 1993.

MENEZES, A. J., E. A. ***Handbook of applied cryptography.*** CRC Pressl., 1996.

MURPHY, S. **The Power of NIST's Statistical Testing of AES Candidates**. *Information Security Group, Royal Holloway, University of London, Egham, Surrey TW20 0EX, U.K.*, 2000.

NAGIREDDY, S. **A pattern recognition approach to block cipher identification.** Dissertação de Mestrado, Department of Computer Science and Engineering Indian Institute of Technology Madras., 2008.

PRADO, P.P.L., C. A. A. M. **Algoritmos para Reconhecimento de Padrões.** *Departamento de Engenharia Elétrica - Universidade de Taubaté*, 2008.

RANDALL, K. e PANOS, L. ***Wireless Security: Models, Threats, and Solutions.*** McGraw - Hill Professional., 2002.

RIVEST, L. **"The RC5 encryption algorithm"**. *Fast Software Encryption, pp.86-96*, 1995.

SINGH, S. ***O livro dos códigos. A ciência do sigilo - do antigo Egito à criptografia quântica***. *Sétima edição*. Record, 2008.

SOTO, R. e LAWRENCE, B. **Randomness Testing of the Advanced Encryption Standard Finalist Candidates**. Technical report, National Institute of Standards and Technology, 100 Bureau Drive, Stop 8930, Gaithersburg, MD 20899-8930, 2000.

SOUZA, W. A. R. **Identificação de Padrões em criptogramas usando técnicas de classificação de textos**. Dissertação de Mestrado, Instituto Militar de Engenharia, 2007.

SPILLMAN, R. **Cryptanalysis of knapsack ciphers using genetic algorithms.** *Cryptologia, 17(4), 367-377.*, 1993.

STALLINGS, W. ***Criptografia e Segurança de Redes*** . Pearson, Quarta edição, 2008.

STINSON, D. ***Cryptography, Theory and Pratice***. Chapman & Hall / CRC, third edition, 2006.

THEODORIDIS, S. e KOUTROUMBAS, K. ***Pattern Recognition.*** Elsivier. Fourth Edition, 2009.

TORRES, R., OLIVEIRA, G., XEXÉO, J., SOUZA, W. e LINDEN, R. **Identificação de chaves e algoritmos criptográficos utilizando Algoritmo Genético e Teoria dos Grafos.** *International Information and Telecommunication Technologies Symposium- 2010. Rio de Janeiro-Brasil*, 2010.

TOU, J. e GONZALEZ, R. C. ***Pattern Recognition Principles.*** Addison-Wesley Publishing Company, Massachusetts., 1981.

UEDA, E. T. e TERADA, R. **A new version of the RC6 algorithm, stronger against χ^2 cryptanalysis**. *Dept. of Computer Science University of São Paulo, Brazil*, 2007.

YASEEN, I. e SAHASRABUDDHE, H. **A genetic algorithm for the cryptanalysis of Chor-Rivest knapsack public key cryptosystem (PKC).** Em *In Proceedings of Third International Conference on Computational Intelligence and Multimedia Applications.*, págs. 81–85, 1999.