MARINHA DO BRASIL
DIRETORIA DE ENSINO DA MARINHA
CENTRO DE INSTRUÇÃO ALMIRANTE WANDENKOLK

CURSO DE APERFEIÇOAMENTO AVANÇADO EM SEGURANÇA DA INFORMAÇÃO E COMUNICAÇÕES

1ºTen (QC-CA) GABRIEL DE CARVALHO ABI-ABIB

UM ESTUDO DE FERRAMENTAS DE CRIPTOANÁLISE BASEADAS EM TÉCNICAS NÃO CONVENCIONAIS

Rio de Janeiro
2020

1ºTen (QC-CA) GABRIEL DE CARVALHO ABI-ABIB

UM ESTUDO DE FERRAMENTAS DE CRIPTOANÁLISE BASEADAS EM TÉCNICAS NÃO CONVENCIONAIS

Monografia apresentada ao Centro de Instrução Almirante Wandenkolk como requisito parcial à conclusão do Curso de Aperfeiçoamento Avançado em Segurança da Informação e Comunicações.

Orientador:
CF (RM1-EN) William Augusto Rodrigues de Souza

CIAW
Rio de Janeiro
2020

ABI-ABIB, Gabriel de Carvalho.
Um estudo de ferramentas de criptoanálise baseadas em técnicas não convencionais / Gabriel de Carvalho Abi-Abib. Rio de Janeiro: CIAW, 2020.
Total de folhas. 42f.: il.

Orientador técnico e acadêmico: CF (RM1-T) William Augusto Rodrigues de Souza.

Monografia (Curso de Aperfeiçoamento Avançado em Segurança da Informação e Comunicações) - Centro de Instrução Almirante Wandenkolk. Centro de Pós-Graduação, 2020.

1. Criptoanálise. 2. Criptografia. 3. Recuperação da Informação. 4. Redes Neurais. I. Centro de Instrução Almirante Wandenkolk. Centro de Pós-Graduação. II. Título.

1ºTen (QC-CA) GABRIEL DE CARVALHO ABI-ABIB

UM ESTUDO DE FERRAMENTAS DE CRIPTOANÁLISE BASEADAS EM TÉCNICAS NÃO CONVENCIONAIS

Monografia apresentada ao Centro de Instrução Almirante Wandenkolk como requisito parcial à conclusão do Curso de Aperfeiçoamento Avançado em Segurança da Informação e Comunicações.

Aprovada em 23 de abril de 2020.

Banca Examinadora:

CMG (RM1-EN) Gian Karlo Huback Macedo de Almeida – CIAW

CF (RM1-T) William Augusto Rodrigues de Souza

Ph. D. Anderson Oliveira da Silva – PUC Rio

CIAW

Rio de Janeiro

2020

Dedico esse trabalho à pessoa que mais me amou na vida, a quem eu devo tudo que sou e o que tenho e que me deixou com uma ferida eterna de saudade no coração: minha mãe.

AGRADECIMENTOS

À minha esposa, Marianny, o sentido da minha vida e motivo da minha alegria de cada dia.

À minha mãe, a quem devo tudo que sou.

Aos meus colegas de turma do CApA-SIC, que me deram força ao longo do curso.

Ao CMG (RM1-EN) Huback, coordenador do curso do CApA-SIC, por todo auxílio e suporte ao longo do curso.

Ao CF (RM1-EN) William, meu orientador, pela prontidão, pelas dicas e por todo encaminhamento deste trabalho.

Aos professores da PUC-RIO, que com empenho e dedicação, conseguiram passar toda a base de conhecimento em Segurança da Informação e Comunicações.

“Em algum lugar, algo incrível está esperando para ser descoberto.”
(Carl Sagan)

UM ESTUDO DE FERRAMENTAS DE CRIPTOANÁLISE BASEADAS EM TÉCNICAS NÃO CONVENCIONAIS

Resumo

Este trabalho é resultado de uma pesquisa exploratória de metodologias de criptoanálise não convencionais. Primeiramente, são apresentadas técnicas estatística (método do χ^2), analítica (função limiar) e de aprendizado de máquina (máquina de vetor de suporte). Posteriormente, são demonstrados métodos baseados em Recuperação da Informação, que são metodologias de agrupamento e classificação de padrões baseadas em processamento de textos. Dessa forma, os criptogramas, cujas cifras necessitam ser identificadas, são interpretados como documentos textuais.

As técnicas de Recuperação da Informação baseadas em Rede Neural artificial e Algoritmo Genético apresentaram bons resultados na identificação dos cinco algoritmos utilizados para gerar os criptogramas. As metodologias são ampliáveis para testes com outras cifras, bem como testes com uma quantidade maior de cifras.

Palavras- chave: Criptoanálise, Recuperação da Informação,

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

TI	Tecnologia da Informação
NIST	*National Institute of Standards and Technology*
ECB	*Electronic Code Book*
CBC	*Cipher Block Chaining*
VI	Vetor de inicialização
DES	*Data Encryption Standard*
IDEA	*International Data Encryption Algorithm*
AES	*Advanced Encryption Standard*
SVM	*Support Vector Machine*
RI	Recuperação da Informação

SUMÁRIO

1 INTRODUÇÃO

1.1 APRESENTAÇÃO

A criptologia é a ciência que abrange a criptografia e a criptoanálise. A criptografia é o conjunto de técnicas voltadas para cifrar informações, permitindo acesso apenas a pessoas autorizadas. Enquanto que a criptoanálise é o conjunto de técnicas com o objetivo de encontrar vulnerabilidades presentes nas cifras (RUKHIN, 1999). A Doutrina de Tecnologia da Informação da Marinha do Brasil define a criptoanálise como sendo *parte da criptologia que estuda as técnicas matemáticas utilizadas para "quebrar" algoritmos criptográficos, buscando vulnerabilidades e falhas de concepção que permitam a obtenção de atalhos na decifração de um texto cifrado, sem o conhecimento da chave empregada*. (BRASIL, 2007).

1.2 JUSTIFICATIVA

O estudo de técnicas de criptoanálise modernas e não convencionais justifica-se pelo fato de que não há forma de garantir absolutamente a segurança de algoritmos criptográficos, por mais sofisticados que sejam. Em outras palavras, um sistema criptográfico é considerado confiável se atender um conjunto de requisitos, porém os testes de segurança são baseados apenas em reprovação (RUKHIN, 1999).

Além disso, diversos trabalhos já identificaram padrões em criptogramas gerados por algoritmos criptográficos, que foram submetidos aos testes do NIST (*National Institute of Standards and Technology*) (CARVALHO, 2006). Tais resultados reforçam a tese sobre a presença de "assinaturas"nos criptogramas, sejam elas geradas pelas chaves ou pelo próprio algoritmo (OLIVEIRA; XEXéO, 2013).

A inviabilidade de se comprovar e garantir a segurança absoluta de um sistema criptográfico abre as portas para o campo da criptoanálise, que tem avançado cada vez mais adentro de técnicas modernas, tais como algoritmos baseados em análises estatísticas, recuperação da informação, redes neurais e algoritmos genéticos, como será apresentado neste trabalho.

1.3 RELEVÂNCIA

A relevância desse trabalho para a Marinha do Brasil fica evidenciada ao se realizar uma análise histórica sobre a importância da criptoanálise durante as duas grandes guerras e em demais conflitos de grande escala. A criptoanálise mostrou-se como uma das principais forças de uma nação em uma guerra, tendo um papel decisivo na história dos conflitos armados, alterando o rumo das batalhas e contribuindo estrategicamente para o sucesso dos vitoriosos (STEVENSON, 2016).

A criptologia deve ser tratada como ponto fundamental em assuntos de defesa nacional em diversos países, inclusive no Brasil. A Doutrina de Tecnologia da Informação da Marinha apresenta como um de seus propósitos "promover atividades de desenvolvimento de criptoanálise operacional, relativas a inteligência e decifragem, e criptoanálise certificacional, relativas a certificação, homologação e verificação da robustez das cifras utilizadas"(BRASIL, 2007).

Este trabalho destina-se a fornecer subsídios à Marinha do Brasil, no que diz respeito à avaliação da viabilidade de uso e desenvolvimento de ferramentes modernas de criptoanálise, a fim de contribuir para o cumprimento dos propósitos de sua Doutrina de TI.

1.4 OBJETIVOS

O propósito primordial deste trabalho é fornecer subsídios à Marinha do Brasil quanto ao emprego de técnicas de criptoanálise não-convencionais, descritas na literatura, a fim de contribuir para o cumprimento de sua Doutrina de TI.

1.4.1 Objetivo geral

Serão apresentadas técnicas que buscam encontrar padrões em coleções de criptogramas, de modo a poder identificá-los e classificá-los. Os métodos apresentados neste trabalho baseiam-se em análises estatísticas, modelagem em espaços vetoriais e em recuperação da informação.

1.4.2 Objetivos específicos

O trabalho tem como objetivos específicos:

a) apresentar alguns dos algoritmos criptográficos mais utilizados (SOUZA, 2011);

b) contextualizar a importância da criptoanálise para segurança das informações e, principalmente, no âmbito militar naval, no qual se insere as atividades da Marinha do Brasil;

c) apresentar os algoritmos de criptoanálise mais utilizados (SOUZA, 2011);

d) apresentar técnicas estatísticas para identificação de cifras, como os métodos do χ^2 e da função limiar;

e) apresentar técnicas baseadas em máquinas de vetor de suporte para classificação de cifras;

f) apresentar as etapas relativas aos sistemas de recuperação da informação;

g) apresentar técnicas de agrupamento e classificação de cifras baseadas em redes neurais e algoritmo genético.

1.5 METODOLOGIA

A apresentação das técnicas de criptoanálise serão realizadas a partir de pesquisa exploratória qualitativa e quantitativa, cuja fonte será livros, teses, dissertações e artigos científicos relativos aos temas: segurança da informação, recuperação da informação e criptologia.

1.5.1 Limites e escopo

A apresentação dos algoritmos de criptografia neste trabalho será limitada a uma visão geral de cada um, contendo apenas as operações principais e características básicas das fases de processamento, visto que o escopo do trabalho é voltado para o campo da criptoanálise.

As técnicas de criptoanálise serão abordadas de modo mais minucioso, incluindo todas as principais equações do modelamento matemático necessário.

O escopo do trabalho limita-se a uma abordagem exploratória de outros trabalhos presentes na literatura, logo não foram realizados testes e experimentos.

1.6 VISÃO GERAL DO TRABALHO

No capítulo 2, é realizada a fundamentação teórica, relativa aos campos da criptografia e da criptoanálise. O capítulo 3 apresenta os seguintes métodos de reconhecimento de cifras: método do χ^2, função limiar e máquinas de vetor de suporte. O capítulo 4 aborda o tema referente à recuperação da informação, apresentando as técnicas baseadas em rede neural e algoritmo genético. O capítulo 5 contém a conclusão do trabalho.

2 FUNDAMENTOS TEÓRICOS

Neste capítulo, primeiramente, serão apresentados os fundamentos básicos relativos à criptografia, sua classificação e seus algoritmos mais relevantes. Em seguida, será apresentada a relevância histórica da criptoanálise no âmbito militar-naval, bem como os métodos mais conhecidos de quebra de cifras, atualmente.

2.1 CRIPTOGRAFIA

A criptografia é a prática de codificar dados através de um algoritmo, de modo que eles não tenham mais o formato original, e assim, não possam mais ser lidos de forma clara. As técnicas de codificação são uma parte importante da segurança da informação, pois protegem os dados confidenciais de ameaças como *malwares* e acesso não autorizado por terceiros. Além disso, a criptografia é uma solução de segurança versátil, pois pode ser aplicada a um dado específico, ou a toda base de dados de um sistema de armazenamento. Os algoritmos criptográficos transformam textos claros em criptogramas, utilizando diversas rodadas de operações e transformações lineares ou não-lineares (CARVALHO, 2006).

Em geral, a criptografia é empregada quando deseja-se garantir um ou mais dos seguintes requisitos de segurança das informações (CARVALHO, 2006):

a) confidencialidade da mensagem: apenas o destinatário deve ser capaz de acessar o conteúdo da mensagem em texto claro;

b) integridade: capacidade do destinatário de verificar se a mensagem foi alterada durante o processo de transmissão;

c) autenticação: capacidade do destinatário de verificar a identidade do remetente;

d) não-repúdio: impossibilidade do remetente negar a autoria de uma mensagem.

Cabe ressaltar que nem todos os sistemas criptográficos são implementados para cumprirem todos os requisitos descritos acima, visto que nem todos os são práticos ou nem mesmo são desejáveis (CARVALHO, 2006).

A criptografia pode ser classificada quanto ao tipo de gerenciamento de chaves utilizadas para os processos de cifragem e decifragem, sendo definida como simétrica quando ambos os processos utilizam a mesma chave ou quando uma das chaves é facilmente obtida a partir da outra (CRYPTO, 2020). A grande vantagem da criptografia simétrica é a sua velocidade, visto que, em geral, os algoritmos são simples e eficientes. Por outro lado, exige um canal seguro para o compartilhamento da chave, o que representa um problema de complexidade crescente, visto que para n usuários, é necessário distribuir $(n(n-1)/2)$ chaves de forma segura. Além disso, não permite a autenticação do remetente e não atende ao requisito de não-repúdio (CRYPTO, 2020). A figura 1 ilustra um cenário de comunicação, que utiliza criptografia simétrica.

Na criptografia assimétrica, cada usuário possui um par de chaves pública e privada

Figura 1 – Comunicação com Criptografia Simétrica (CRYPTO, 2020).

(CRYPTO, 2020). Uma mensagem é criptografada com a chave pública do destinatário, que por sua vez, irá decriptá-la, usando sua chave privada. Esse processo fornece muito mais segurança à comunicação, visto que não há necessidade de envio da chave para o destinatário, como na criptografia simétrica. A desvantagem advém do fato de, por serem mais complexos, os algoritmos da criptografia assimétrica são mais lentos e necessitam de mais poder de processamento (CRYPTO, 2020). A figura 2 ilustra o processo de comunicação com criptografia assimétrica.

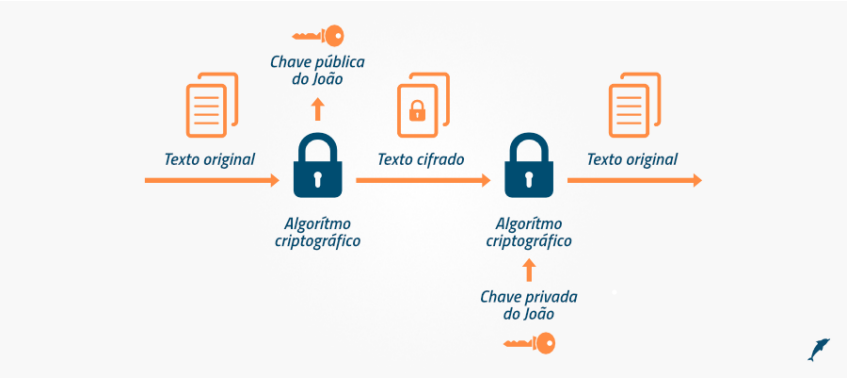

Figura 2 – Comunicação com Criptografia Assimétrica (CRYPTO, 2020).

O algoritmo de criptografia assimétrica mais famoso é o RSA, criado em 1977 por Ron Rivest, Adi Shamir, e Leonard Adleman (CRYPTO, 2020). O RSA calcula as chaves pública e privada a partir do produto N de dois números primos aleatórios muito grandes, com 2048 *bits* cada (CRYPTO, 2020). Esse processo garante a segurança do RSA, visto que para quebrar a chave privada de um usuário, seria necessário fatorar o número N, de 4096 *bits*, o que é computacionalmente inviável.

Quanto ao modo de processamento, a criptografia também pode ser classificada como sendo cifragem de fluxo ou em blocos. A cifragem de fluxo ou cadeia combina, através de uma operação *XOR*, cada *bit* ou *byte* da mensagem original com cada *bit* ou *byte* de uma chave advinda de um gerador de dígitos pseudo-aleatório (CRYPTO, 2020).

Já na cifragem em blocos, a mensagem original é dividida em blocos de tamanhos, que dependem do algoritmo utilizado e, em alguns casos, do tamanho da chave. Um dos méto-

dos de cifragem em bloco é o ECB (*Electronic Code Book*). Nesse caso, os blocos são cifrados e concatenados na mesma ordem do texto original. O problema fundamental deste método reside no fato de que, se for utilizada a mesma chave, blocos legíveis idênticos resultarão em blocos cifrados idênticos, podendo, assim, revelar padrões contidos no texto plano. A figura 3 apresenta o processo de cifragem com método ECB (RIBEIRO; ROIHA, 2010).

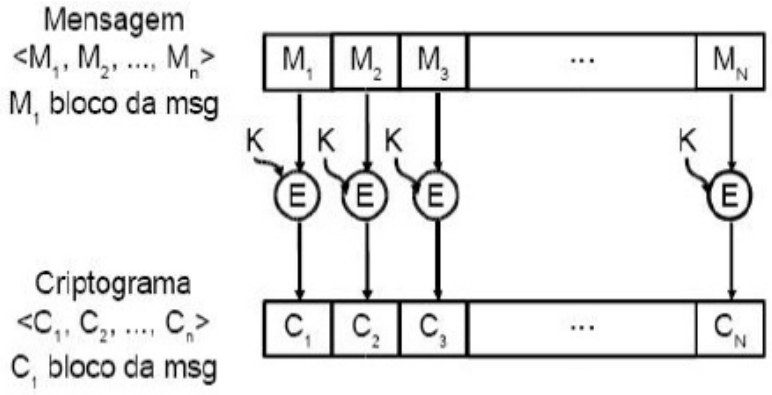

Figura 3 – Processo de cifragem com EBC (RIBEIRO; ROIHA, 2010).

Outro método de cifragem em bloco é o CBC (*Cipher Block Chaining*). Nesse modo, a operação *XOR*é realizada entre um bloco de texto plano com o bloco cifrado antecedente, eliminando, assim, o problema de blocos idênticos gerarem blocos cifrados idênticos. Além da chave, faz-se necessária a utilização de um Vetor de inicialização (VI), já que não há bloco cifrado para o bloco inicial (RIBEIRO; ROIHA, 2010). A figura 4 ilustra o processo de cifragem, através do método CBC.

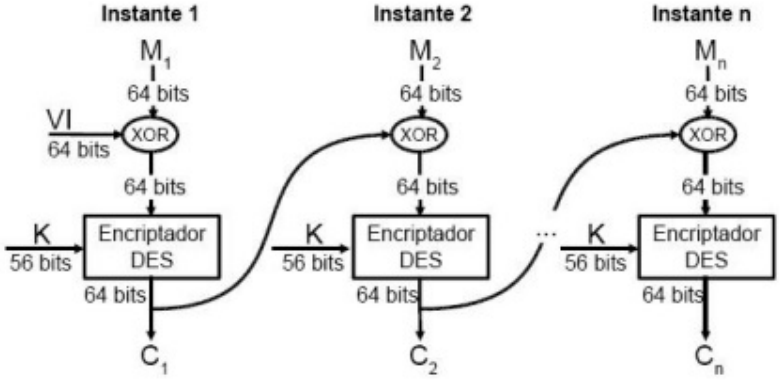

Figura 4 – Processo de cifragem com CBC (RIBEIRO; ROIHA, 2010).

O processo de cifragem com CBC ilustrado na figura 4 utiliza para a encriptação dos blocos o algoritmo DES (*Data Encryption Standard*) (NIST, 1999). O DES, criado pela IBM, foi definido como o primeiro algoritmo padrão de criptografia simétrica com cifragem em blocos. Atualmente, é considerado inseguro, para a maior parte das aplicações. Isso deve-se, principalmente, ao tamanho pequeno da chave de 64 *bits*, sendo que apenas 56 *bits* são de fato

utilizados para as operações de encriptação. Uma versão mais segura do DES é o 3DES que usa três chaves de 64 *bits*, porém apenas 56 *bits* de cada chave são efetivamente utilizados no processo. No 3DES, os dados são encriptados pela primeira chave, decriptados pela segunda chave e, novamente encriptados pela terceira (BIHAM; SHAMIR, 1991).

O DES é uma cifra que baseia-se em uma rede de Feistel. Esse tipo de estrutura de cifras utiliza operações de cifragem e decifragem semelhantes. A única diferença entre os dois processos é a ordem das subchaves utilizadas (BIHAM; SHAMIR, 1991).

A operação básica de uma rede de Feistel separa um bloco de texto plano em duas metades, L_0 e R_0. Seja f a função rodada e $k_0, k_1, ..., k_n$ as subchaves para as rodadas $i = 0, 1, ..., n$, respectivamente. Dessa forma, para cada rodada i, tem-se (BIHAM; SHAMIR, 1991):

i. $L_{i+1} = Ri$
ii. $Ri + 1 = Li \oplus f(R_i, K_i)$

Para a decifração, as subchaves devem ser computadas na ordem inversa, com $i = n, n - 1, ..., 0$, porém as operações são as mesmas (BIHAM; SHAMIR, 1991):

i. $R_i = L_{i+1}$
ii. $Li = Ri + 1 \oplus f(L_{i+1}, K_i)$

Uma cifra em bloco com uma estrutura semelhante ao DES, porém com chave simétrica de 128 *bits*, surgiu em 1991 com o nome de IDEA (*International Data Encryption Algorithm*), criada por Xuejia Lai e James Massey. O IDEA é implementado usando três grupos algébricos cujas operações são misturadas. As operações são *XOR*, adição em módulo 2^{16} e multiplicação em módulo $2^{16} + 1$ (BIHAM; SHAMIR, 1991).

Em 2001, o NIST elegeu, através de concurso um novo algoritmo padrão criptográfico, com intuito de substituir o DES pelo AES (*Advanced Encryption Standard*). O vencedor do concurso foi o algoritmo Rijndael, desenvolvido por John Daemen e Vincent Rijmen, que utiliza blocos de 128 *bits* e chaves de 128, 192 ou 256 *bits* (NIST, 2001). O AES foi amplamente difundido e é utilizado até hoje em diversas aplicações, como sendo, por exemplo, recurso criptográfico do protocolo de segurança de acesso à rede sem fio WPA2.

Além do algoritmo vencedor de Rijndael, outros quatro algoritmos foram os finalistas do concurso do NIST para eleger o AES: RC6 (RSA Laboratories), MARS (IBM), Serpent (Ross Anderson, Eli Biham e Lars Knudsen) e o TwoFish (Bruce Schneier, John kelsey e outros) (SOUZA, 2011). A seguir será apresentada uma breve explanação sobre cada um dos finalistas.

2.1.1 RC6

O RC6 é uma cifra de blocos, baseada em sua versão anterior, RC5, porém modificada para atender aos requisitos do concurso que elegeu o AES. Trata-se de uma cifra de chave simétrica, que varia de 16 a 255 *bytes* e opera em blocos de 128 *bits*. Para o concurso, foi implementado com 20 iterações, porém pode ser parametrizado para executar entre 8 e 20 iterações (SOUZA, 2011).

A cifra é parametrizada por 3 parâmetros: w, que representa o tamanho do bloco, r é número de rodadas e b , que é o tamanho da chave em *bytes*. A partir da chave são derivadas $2r + 4$ chaves de tamanho w, utilizadas no processo de cifragem, que é realizado por meios das seguintes operações: *XOR*, *ADD*, Rotação à esquerda, Rotação à direita, multiplicação, permutação e Subtração (SOUZA, 2011).

2.1.2 MARS

A cifra MARS é um dos algoritmos finalistas do concurso que elegeu o AES que baseia-se em rede de Feistel. Opera com blocos de 128 *bits* e chaves que podem variar entre 128 e 400 *bits* e realiza transformações dos tipos substituição, *XOR*, *ADD*, rotação, expansão, multiplicação e subtração. O algoritmo constitui-se de três fases, sendo a primeira com oito rodadas, a segunda com 16 rodadas, sendo 8 em modo *forward* e 8 em modo *backward* e a terceira com mais rodadas (SOUZA, 2011). A figura 5 ilustra o processo.

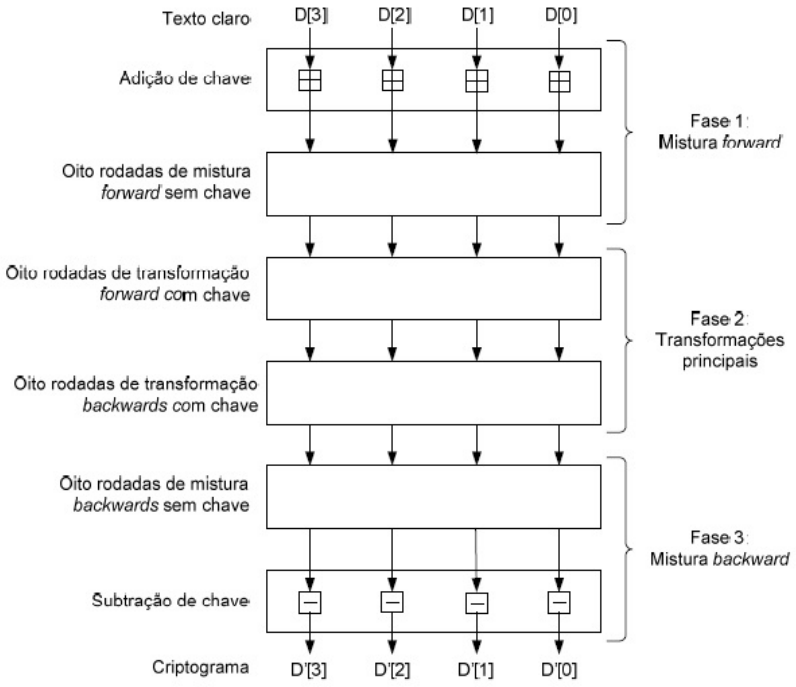

Figura 5 – Algoritmo de Cifragem MARS (SOUZA, 2011).

2.1.3 Serpent

O Serpent é um algoritmo de blocos que opera com 128 *bits* e chave simétrica de 128, 192 ou 256 *bits*. Trata-se de uma cifra com estrutura em rede de Feistel de 32 rodadas e utiliza operações como *XOR*, rotação, permutação, substituição e deslocamento (SOUZA, 2011).

O processo de cifragem constitui de três fases. A primeira fase consiste em uma permutação aplicada ao bloco em texto claro. Na segunda fase, ocorrem as 32 rodadas, nas quais são realizadas as operações de inserção da chave criptográfica, substituição e uma transformação linear, que inclui as operações de rotação, deslocamento e *XOR*. Na terceira fase, após as 32 rodadas, ocorre mais uma permutação (SOUZA, 2011).

2.1.4 Twofish

O Twofish também é uma cifra de bloco de 128 *bits*, com chave simétrica de 128, 192 ou 256 *bits*. Trata-se também de uma rede de Feistel, porém com 16 rodadas e nao 32 como no Serpent. O processo de cifragem é mais complexo, pois opera com estruturas específicas, tais como, *Pseudo-Hadamard Transforms* (PHT), Matrizes MDS e *Whitening* (SOUZA, 2011).

Em suma, o processo divide-se em três fases. Na primeira fase, é realizado um *XOR* entre quatro sub-blocos da mensagem com quatro sub-blocos das chaves parciais. Na segunda fase, ocorrem as 16 rodadas, nas quais são realizadas as operações com as estruturas citadas acima, além das operações de *ADD*, *XOR*, rotação à esquerda, rotação à direita, multiplicação e substituição. Na terceira fase, ocorrem uma permutação, que desfaz a permutação realizada pela rede de Feistel e um *XOR* entre os quatro blocos resultantes das 16 rodadas com quatro sub-blocas de chaves parciais (SOUZA, 2011).

2.1.5 Rijndael

O vencedor do concurso que elegeu o AES, o algoritmo Rijndael opera em blocos de 128 *bits*, com 10, 12 ou 14 rodadas, para chaves simétricas com tamanho de 128, 192 e 256 *bits*, respectivamente. A estrutura da cifra não se baseia em uma rede de Feistel e realiza as seguintes operações descritas a seguir (SOUZA, 2011):

a) *ADDRoundKey*: trata-se se de uma operação *XOR* realizada *bit* a *bit* entre o bloco da mensagem e a chave;

b) *SubBytes*: trata-se de um cruzamento realizado entre uma matriz (*S-box*) e os *bytes* do bloco de entrada;

c) *ShiftRows*: trata-se de uma operação de permutação entre os *bytes* de um bloco;

d) *MixColumns*: nessa operação, cada bloco de 128 *bits* é dividido em quatro sub-blocos de 32 *bits*. Cada sub-bloco é multiplicado por uma matriz quadrada de 32x32 *bits*.

e) *KeyExpansion*: trata-se de uma operação para gerar $nr + 1$ subchaves a partir da chave criptográfica, sendo que nr é o número de rodadas. Cada uma dessas subchaves será

utilizada para cada operação *ADDRoundKey*.

f) *RotWord*: Consiste em uma operação de deslocamento cíclico de um sub-bloco *bytes* à esquerda.

A figura 6 ilustra o processo das operações da cifra Rijndael, com $nr = 10$ e o valor inicial de i igual a 1.

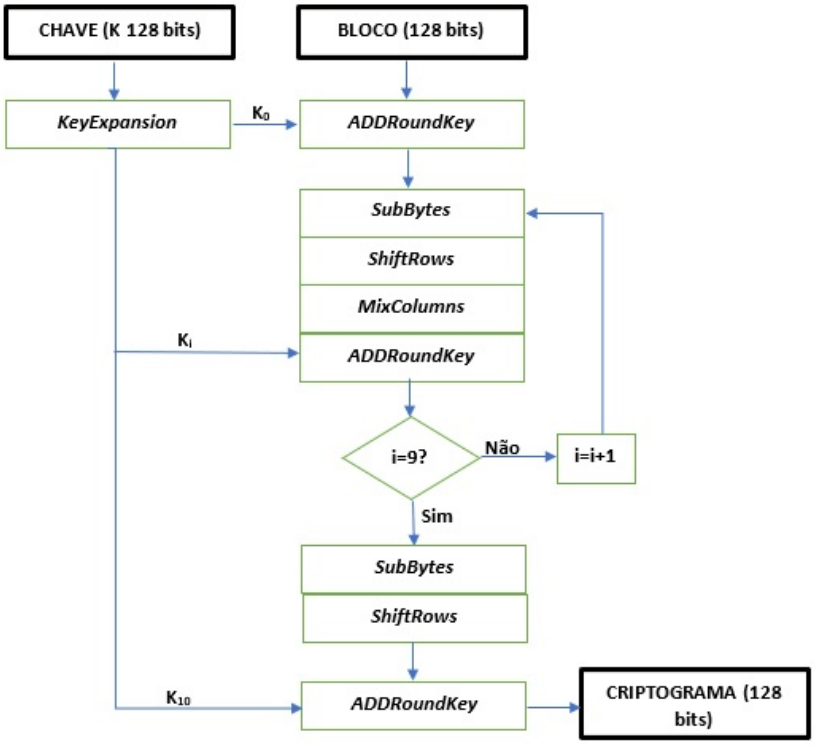

Figura 6 – Cifra de Rijndael com $nr = 10$. Adaptado de LAMBERT (2004).

2.2 CRIPTOANÁLISE

A criptoanálise não se resume simplesmente como um conjunto de técnicas com o objetivo de decifrar mensagens ou desvendar chaves criptográficas. Na verdade, a criptoanálise busca encontrar vulnerabilidades de uma cifra, com um esforço computacional menor do que seria exigido através da força bruta, evidenciando que a mesma não funciona como o esperado. Trata-se de um ramo da criptologia que caminha a serviço da criptografia, a fim de testar e garantir a robustez das cifras (CARVALHO, 2006).

O processo da criptoanálise também pode ser entendido como ataque, que por sua vez, pode ser classificado quanto ao conjunto de informações disponíveis da seguinte forma: Somente com texto cifrado disponível, com texto plano e cifrado disponíveis, com texto plano escolhido dentre vários e com texto claro modificado com base em resultados anteriores (CARVALHO, 2006).

2.2.1 Criptoanálise no âmbito militar-naval

Durante a Primeira Guerra Mundial, a Inglaterra mantinha um escritório, conhecido como *Sala 40*, onde concentravam-se os maiores especialistas em decifração, comandados pelo Almirante William Reginald Hall. A *Sala 40* desempenhou papel decisivo em diversos combates navais durante a guerra. Através da detecção de missões alemães no Mar do Norte, frotas britânicas foram enviadas, a fim de interceptá-las, desencadeando batalhas como as de *Dogger Bank* e *Jutlândia* (STEVENSON, 2016).

Outra importante e decisiva contribuição da *Sala 40* foi a interceptação e decodificação do Telegrama *Zimmermman*, despachado pelo ministro do exterior alemão para o embaixador alemão no México. O telegrama ordenava uma aproximação do embaixador com o governo mexicano, com o intuito de formar uma aliança contra os Estados Unidos. O texto interceptado e decifrado pelos ingleses apressou a entrada dos norte-americanos na guerra (STEVENSON, 2016).

Segundo Alvarenga (2010), até o final da primeira guerra mundial, a inteligência norte-americana pouco havia desenvolvido no que diz respeito ao trabalho de cifragem e decifragem. Em 1917, com a criação do Departamento de Cifras do Exército, cuja missão inicial era quebrar códigos de mensagens interceptadas dos alemães, o ramo da criptoanálise começou a se desenvolver. A partir de 1920, houve um avanço substancial com o desenvolvimento de métodos estatísticos aplicados à criptoanálise.

Paralelamente, desde 1918, o Serviço de Inteligência da Marinha norte-americana era responsável por acompanhar e interceptar os sistemas de comunicação japonesa, em especial os códigos utilizados pela Marinha nipônica. Em 1931, a Marinha norte-americana dedicou-se inteiramente em tentar decifrar dois grandes e importantes documentos japoneses: o *Código Diplomático* e as *Comunicações Navais*. Para isso, instalou várias bases de escuta e interceptação na Ásia (ALVARENGA, 2010).

Já durante a Segunda Guerra Mundial, criptógrafos da Marinha dos Estados Unidos quebraram o sistema JN-25 de criptografia japonesa, resultando na vitória norte-americana da batalha de *Midway* (ALVARENGA, 2010). Ainda antes da guerra começar, os Estados Unidos já tinham quebrado um dos sistemas mais seguros de codificação diplomática japonês, a *Purple*, nome dado pelo americanos à máquina eletromecânica japonesa de codificação de mensagens.

Desde a Primeira Guerra, o uso de codificação polialfabética através de máquinas mecânicas e eletromecânicas já era muito frequente. Os alemães fizeram uso constante de várias versões de um rotor eletromecânico, denominado *Enigma*, inclusive durante a Segunda Guerra, para enviar mensagens contendo dados de posicionamento dos navios de sua frota naval. A quebra da cifra da *Enigma* iniciou-se com a dedução da sua estrutura detalhada por um escritório polonês, em 1932, culminando com a sua decodificação pelos britânicos, fato relevante para a vitória dos aliados na guerra (STEVENSON, 2016).

Todos estes fatos são apenas alguns exemplos da importância estratégica da criptoanálise no âmbito militar-naval. Ainda que, atualmente, os sistemas criptográficos estejam cada

vez mais confiáveis, não há forma de garantir de forma absoluta a segurança das cifras.

2.2.2 Criptoanálise Linear e Diferencial

Os métodos mais conhecidos de quebra de cifras são a criptoanálise linear e a criptoanálise diferencial. A primeira foi desenvolvida em 1993 como tentativa de ataque ao DES (MATSUI,). Em suma, a criptoanálise linear analisa as relações estatísticas entre os *bits* de mensagens em texto plano e os *bits* das cifras correspondentes. O objetivo é tentar predizer os valores dos *bits* da chave, através de aproximações lineares, quando são conhecidas algumas mensagens em texto plano e seus respectivos criptogramas. Em seu trabalho, MATSUI (1993) enuncia que a criptoanálise linear busca encontrar a expressão linear efetiva para um dado algoritmo, de acordo como descrito na equação 2.1:

$$P[i_1, i_2, ..., i_a] \oplus C[j_1, j_2, ..., j_b] = K[k_1, k_2, ..., k_c] \tag{2.1}$$

onde i_a, j_b e k_c representam, respectivamente, as posições fixas dos *bits* do texto plano, do texto cifrado e da chave. A partir de textos claros escolhidos aleatoriamente e seus criptogramas correspondentes, a probabilidade *p* da equação ser válida deve ser diferente de $\frac{1}{2}$. A magnitude de *p* é dada por $\mid p - \frac{1}{2} \mid$ e define a efetividade da expressão linear.

Outro método clássico é a criptoanálise diferencial (BIHAM; SHAMIR, 1991), que baseia o seu ataque comparando, iterativamente, a diferença entre pares de criptogramas conhecidos com a diferença entre as mensagens em texto claro correspondentes. O termo "diferença", nesse caso, deve ser entendida como uma operação *XOR* entre os pares de textos. O melhor ataque realizado contra o DES, utilizando criptoanálise diferencial, com 16 iterações, requer 2^{47} pares de mensagens escolhidas. Enquanto que a criptoanálise linear consegue quebrar DES com 2^{43} pares de texto claro. Ambos os métodos se mostram muito mais eficientes do que a quebra por força bruta, cuja complexidade é de 2^{56}, porém ainda exigem elevado custo computacional.

Além das criptoanálises linear e diferencial, outros tipos de ataques são propostos pelo NIST como necessários, porém não suficientes, para avaliar a robustez de um algoritmo criptográfico (NIST, 2008). A robustez está associada à capacidade do algoritmo funcionar como um gerador de números pseudo-aleatórios, visto que o mesmo deve gerar criptogramas a partir de textos claros (SOUZA, 2011). Dessa forma, encontrar padrões em criptogramas sugere a existência de vulnerabilidades no algoritmo e compromete sua robustez.

Durante as últimas duas décadas, técnicas baseadas em análises estatísticas, recuperação da informação e inteligência computacional foram desenvolvidas e aprimoradas, aperfeiçoando os ataques a algoritmos criptográficos, com enfâse na identificação de cifras e na busca de padrões e assinaturas nos seus respectivos criptogramas (OLIVEIRA; XEXéO, 2013). Os próximos capítulos dedicam-se a apresentar algumas dessas técnicas.

3 RECONHECIMENTO DE CIFRAS

Como ressaltado anteriormente, a quebra de um algoritmo criptográfico não significa necessariamente decifrar mensagens ou desvendar as chaves. Quebrar uma cifra significa encontrar alguma vulnerabilidade no processo criptológico. Uma das metodologias de ataque concentra-se em identificar cifras de blocos, isto significa que o esforço consiste em descobrir o algoritmo criptográfico a partir somente do criptograma gerado pela cifra (SOUZA, 2011). Alguns desses modelos serão apresentados a seguir. Vale ressaltar que os métodos que serão abordados utilizam criptogramas gerados pelo modo completo das cifras, ou seja, as cifras não foram submetidas qualquer processo de enfraquecimento.

3.1 MÉTODO DO χ^2

O método do χ^2 é um cálculo estatístico utilizado para testar se uma distribuição obtida experimentalmente é consistente com uma distribuição esperada. Seja $X = x_1, x_2, ..., x_n$, tal que $x_i \in a_1, a_2, ..., a_{m-1}$, uma variável aleatória com distribuição de probabilidade **p** desconhecida (SOUZA, 2011). A estatística χ^2 de X estima a distância entre a distribuição observada **p** e a distribuição uniforme esperada $\pi = (\pi_0, \pi_1, ...\pi_{m-1})$. O cálculo da estimativa é definida pela equação 3.1, onde $N_{aj}(X)$ representa o número de vezes que X assume o valor a_j (SOUZA, 2011).

$$\chi^2 = \sum_{i=0}^{m-1} \frac{(N_{aj}(X) - n\pi_i)^2}{n\pi_i} \tag{3.1}$$

A estatística do χ^2 foi utilizada primeiramente por Vaudenay (1996) para ataque ao DES e, posteriormente, por Knudsen e Meier (2001) para verificar se um conjunto de criptogramas foi gerado pelo algoritmo RC6. O teste obteve 95% de sucesso para quebrar o RC6 implementado com 15 rodadas. Uma versão modificada do RC6, denominada RC6T, foi proposta por Terada e Ueda (2009), a fim de tornar o algoritmo mais forte contra o ataque do χ^2.

3.2 FUNÇÃO LIMIAR

Maheshwari (2001) explora o conceito de função limiar na tentativa de classificar criptogramas gerados pelas cifras DES ou IDEA. Neste caso, a função limiar é definida como uma função tal que, $f(x) = 1$, para o DES e $f(x) = 0$, para o IDEA. Em suma, o objetivo do método é encontrar um vetor de números reais $c = [c_0, c_1, ..., c_{319}]$ e um número real T que satisfaça:

i. $f = \sum_{t=0}^{319} c_t b_t \geq T$, para o DES; e

ii. $f = \sum_{t=0}^{319} c_t b_t < T$, para o IDEA.

Onde $b = [b_0, b_1, ..., b_{319}]$ são os 320 primeiros *bits* de um criptograma gerado pelas cifras DES ou IDEA.

O método leva em consideração a função $Z = c_0 + c_1 + ... + c_{319}$, que deve ser maximizada, sujeita às seguintes restrições:

i. $-c_0 - c_1 - ... - c_{319} + T$, para o DES; e

ii. $= c_0 + c_1 + ... + c_{319} - T$, para o IDEA.

Os valores de c e T foram encontrados através de técnicas baseadas em programação linear. Cada criptograma foi dividido em *m* segmentos de 320 *bits*. Foi calculada a razão *R* entre a quantidade *n* de segmentos que obtiveram valor da função maior que T e o total *m* de segmentos. Os resultados foram ligeiramente positivos, porém os valores de R ficaram muito próximos tanto para o DES quanto para o IDEA. Maheshwari (2001) utiliza em seu trabalho apenas propriedades lineares, mas ressalta que melhores resultados podem ser obtidos, se técnicas não-lineares forem introduzidas na computação das funções limiar.

Trabalhos subsequentes obtiveram resultados satisfatórios em experimentos de classificação de cifras com função limiar. Chandra e The (2002) definiu quatro modelos de testes (*good-test* estático, *good-test* dinâmico, *good-test* dinâmico extendido e índice de subarquivo) para classificar criptogramas do DES e IDEA, baseando-se na função limiar estabelecia por Maheshwari (2001). O melhor resultado foi obtido com o *good-test* dinâmico com criptogramas de 200 *Kbytes*, que conseguiu identificar 96% dos criptogramas obtidos a partir do DES e 93% dos criptogramas obtidos pelo IDEA.

3.3 MÁQUINAS DE VETOR DE SUPORTE

O conceito de máquina de vetor de suporte (SVM, do inglês: *support vector machine*) está associado a um conjunto de métodos de treinamento e aprendizado usados para classificar dados através da análise de dados e reconhecimento de padrões (ZUBEN, 2003).

Um SVM pode ser entendido como um classificador linear binário não probabilístico. Primeiramente, na etapa de treinamento, amostras são formadas pelo conjunto de dados de entrada relacionadas às suas respostas previamente classificadas (rótulos). Após o treinamento, o objetivo é classificar amostras ainda não rotuladas. As entradas são mapeadas para um espaço de características (*featured space*), a fim de separá-las linearmente em duas classes por uma linha de separação, conhecida como hiperplano. Um hiperplano ótimo é aquele que apresenta máxima margem de separação ρ, como ilustrado na figura 7 (ZUBEN, 2003).

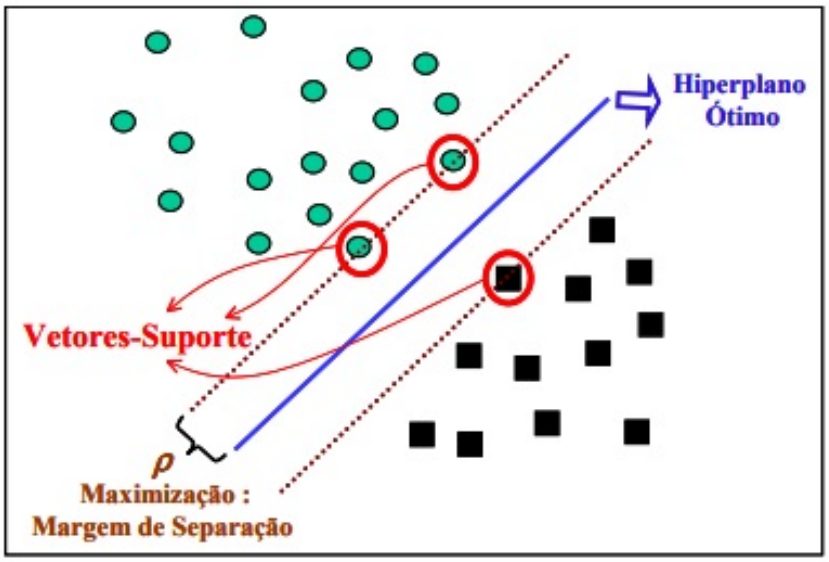

Figura 7 – Hiperplano ótimo para classificação de dados a partir de um SVM (ZUBEN, 2003).

DILEEP e SEKHAR (2006) utilizaram SVM para classificar cifras de bloco DES, Triple DES, Blowfish, AES e RC5. Os criptogramas foram representados como vetores. Os vetores foram compostos através de duas técnicas: a partir de sequências fixas de 16 *bits* que mais se repetiam em um criptograma, ou a partir de sequências variáveis de *bits* delimitadas pelas três sequências que mais se repetiam em um conjunto de 50 criptogramas. Essas bases constituíram um dicionário utilizado para a fase de treinamento da SVM.

Como descrito por Souza (2011), o melhor resultado do trabalho de DILEEP e SEKHAR (2006) foi obtido quando se utilizou uma SVM para cada cifra, dicionários específicos, sequência de tamanho fixo e função gaussiana como núcleo da SVM. As cifras DES, Triple DES e Blowfish, em modo de operação ECB, foram identificadas com 97,78% de precisão. Esse resultado foi obtido com todos os criptogramas gerados pela mesma chave. Vale ressaltar que todos os melhores resultados foram obtidos quando a mesma chave foi utilizada na fase de teste e na fase de classificação. Quando chaves diferentes foram utilizadas para cada fase, a precisão dos teste caiu significativamente.

4 RECUPERAÇÃO DA INFORMAÇÃO

De acordo com WIVES (1997), sistemas de recuperação da informação (RI) possibilitam que usuários encontrem dados que procuram sem precisar analisar todas as informações contidas na base. A fonte de dados podem ser imagens, sons, vídeos e textos. Em suma, os sistemas de RI comparam uma declaração formal, denominada de consulta, com os dados armazenados em uma base de dados (FRAKES, 1997). A aplicação de técnicas de RI na criptoanálise consiste em entender os criptogramas como textos escritos em uma linguagem desconhecida.

As técnicas de RI são técnicas de processamento de textos, que baseiam-se em representações linguísticas associadas às relações sintáticas, semânticas ou discursivas. Porém, para criptogramas, esse processamento fica restrito a uma análise unicamente estrutural, visto que, a priori, não há relações semânticas e/ou sintáticas entre seus elementos (CARVALHO, 2006).

Dessa forma, um criptograma é considerado um documento textual com alfabeto binário e, para o caso de cifras em bloco, a unidade básica é o bloco. De forma geral, um sistema RI busca identificar a distribuição de palavras sem repetição de uma coleção de documentos, para depois compará-los e verificar o grau de similaridade entre eles. Analogamente, quando o documento é um criptograma gerado por uma cifra em bloco, o sistema RI baseia-se na distribuição de blocos dentro de uma coleção de criptogramas (SOUZA, 2011).

4.1 MODELAGEM EM ESPAÇO VETORIAL

Como descrito por Souza (2011), os criptogramas podem ser modelados dentro de um espaço vetorial de dimensão n, sendo n o número de blocos diferentes no conjunto de criptogramas. Nesse modelo, cada bloco é um eixo e cada criptograma é um ponto do espaço vetorial, como ilustrado pela figura 8.

Como observado na figura 8, a distância entre documentos (criptogramas) está associada ao grau de similaridade entre eles. Dessa forma, textos que possuem mais termos (blocos) em comum estão localizados em uma mesma região do espaço vetorial, indicando que são documentos semelhantes (CARVALHO, 2006).

4.2 CÁLCULO DE SIMILARIDADE

Dentre as diversas técnicas de medição de similaridade existentes entre dois documentos, a mais utilizada em recuperação das informações é a do cosseno (HARMAN, 1992), cujo cálculo está descrito na equação 4.1.

$$\cos(\vec{c_i}, \vec{c_j}) = \frac{\sum_{k=1}^{n}(c_{k,i} * c_{k,j})}{\sqrt{\sum_{k=1}^{n}(c_{k,i})^2 \times \sum_{k=1}^{n}(c_{k,j})^2}} \tag{4.1}$$

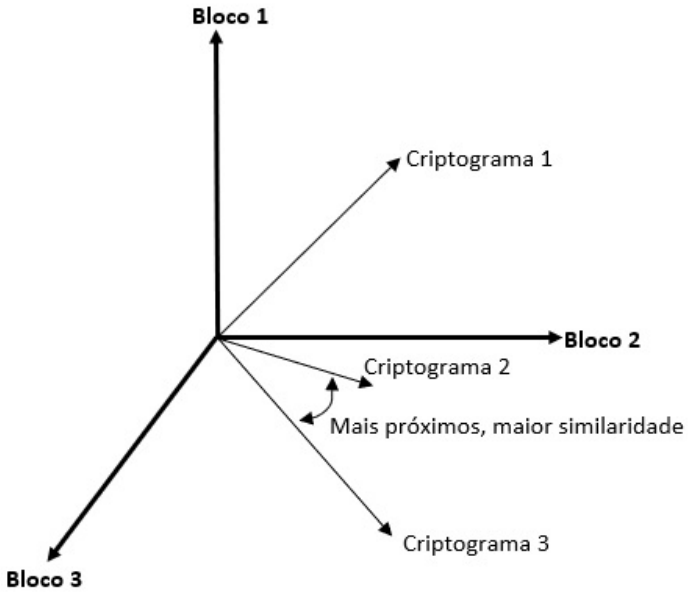

Figura 8 – Modelagem de criptogramas em espaço vetorial. Adaptado de CARVALHO (2006).

Onde $\vec{c_i} = (c_{1,i}, c_{2,i}, ..., c_{n,i})$ e $\vec{c_j} = (c_{1,j}, c_{2,j}, ..., c_{n,j})$ são representações vetoriais dos criptogramas c_i e c_j. O valor de $c_{k,i}$ é a frequência do *k-ésimo* bloco do criptograma c_i e o valor de $c_{k,j}$ é a frequência do *k-ésimo* bloco do criptograma c_j. A partir do cálculo utilizando a equação 4.1, monta-se a matriz de similaridade, que armazena os valores de similaridade entre todos os pares de criptogramas da coleção (SOUZA, 2011).

4.3 AGRUPAMENTO DE CRIPTOGRAMAS

Definido o método de cálculo de similaridade entre criptogramas, a próxima etapa do sistema de recuperação da informação é definir a técnica de agrupamento dos n criptogramas em m grupos. Um dos problemas consiste na dificuldade em determinar o valor de m, visto que, a priori, é desconhecido.

Os métodos mais simples são os não-hierárquicos, que não utilizam sobreposição para realizar a divisão dos n objetos em m grupos. O agrupamento é realizado seguindo critérios relacionados às características dos objetos. São métodos heurísticos, visto que alguns parâmetros são definidos a priori, como número de grupos, tamanho dos grupos e critérios de agrupamento (CARVALHO, 2006).

Métodos hierárquicos são mais complexos, mas não dependem de critérios estabelecidos heuristicamente, possibilitando maior flexibilidade no agrupamento. Os objetos iniciam o processo de forma aninhada e os pares são sucessivamente formados, até que todos os elementos estejam conectados.

Em recuperação das informações, o método hierárquico mais utilizado é o aglome-

rativo, que resulta no final do processo em apenas um único grupo, caso nenhum critério de parada tenha sido estabelecido. O processo pode ser representado graficamente como um dendrograma (RASMUSSEN, 1992), ilustrado na figura 9. Pode-se observar pelo dendrograma que os grupos são formados a partir de um determinado valor de similaridade (SOUZA, 2011).

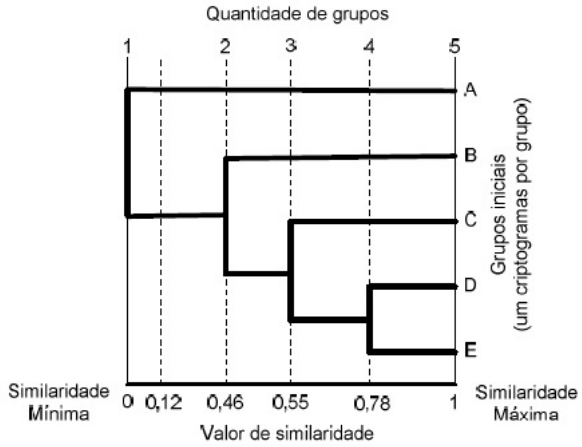

Figura 9 – Dendrograma (SOUZA, 2011).

CARVALHO (2006) descreve o algoritmo de agrupamento aglomerativo da seguinte forma:

a) É formado um grupo para cada um dos n objetos;

b) Calcular a similaridade de todos os pares de grupos, segundo a equação 4.1 e montar a matriz de similaridade;

c) Formar um novo grupo com os dois grupos mais similares;

d) Remontar a matriz de similaridade com o novo grupo e os grupos restantes; e

e) Repetir os passos *c* e *d* até que um critério de parada seja alcançado ou até restar um único grupo.

O passo *d* do algoritmo de agrupamento aglomerativo exige que novas similaridades entre o novo grupo formado e os grupos restantes sejam calculadas, a fim de atualizar a matriz de similaridade. Os métodos mais conhecidos são ligação simples, ligação completa e ligação por média de grupos.

Na ligação simples, a similaridade do novo grupo formado é determinada pela maior das similaridades entre os grupos unidos. Já na ligação completa, a nova similaridade é determinada pela menor das similaridades dos grupos. E na ligação completa, é calculada a média ponderada entre as similaridades dos grupos e o resultado é atribuído ao novo grupo formado.

4.4 AVALIAÇÃO DO AGRUPAMENTO

Souza (2011) utiliza para avaliação da qualidade dos agrupamentos técnicas baseadas em revogação e precisão, medidas definidas nos trabalhos de Yates e Neto. (1999) e Fung (2003).

Como descrito em Souza (2011), seja K , o conjunto de criptogramas cifrados pelo algoritmo Δ (criptogramas relevantes) e seja G, o conjunto de criptogramas agrupados pelo método de recuperação da informação descrito acima. Os valores de precisão e revogação são calculados pelas equações 4.2 e 4.3.

$$R = \frac{n}{k} \tag{4.2}$$

$$R = \frac{n}{g} \tag{4.3}$$

Onde k é quantidade de elementos do conjunto K, g é a quantidade de elementos do conjunto G e n é quantidade de elementos contidos na interseção dos conjuntos K e G, ou seja, n representa a quantidade de criptogramas recuperados que de fato foram gerados pela cifra Δ. A figura 10 ilustra a distribuição dos criptogramas nos conjuntos.

Figura 10 – Criptogramas Relevantes e Recuperados. Adaptado de Souza (2011).

4.5 REDE NEURAL AUTO-ORGANIZÁVEL

Souza (2007) desenvolveu uma rede neural artificial em linguagem Java, para agrupamento e classificação de criptogramas, que utiliza a abordagem estatística dos Sistemas de Recuperação da Informação. A rede neural desenvolvida em Souza (2007) foi aperfeiçoada e testada com maior poder de processamento em Souza (2012).

Redes neurais artificiais têm suas raízes em uma base multidisciplinar, que envolve campos de estudo na matemática, física, computação, neurociência e engenharia. Por ser uma técnica de inteligência artificial, é um método de computação, que distingui-se dos demais métodos apresentados neste trabalho. A sua habilidade de aprendizado a partir de uma etapa de treinamento possibilita uma ampla gama de aplicações, dentre elas o reconhecimento de padrões.

A unidade elementar de uma rede neural é o neurônio. Cada neurônio possui um vetor de pesos sinápticos, que determina como se ligará aos demais. O conceito de auto-organização refere-se à sua capacidade de se modificar em função da sua experiência e de seu relacionamento com o ambiente, sem a supervisão de um agente externo. Dessa forma, a rede neural cria um mapa topográfico, denominado mapa de Kohonen, que vai se modificando com o decorrer das iterações, a partir unicamente das entradas fornecidas. No final do processo, no mapa resultante, entradas similares estarão próximas umas das outras. A figura 11 ilustra uma arquitetura típica de uma rede neural auto-organizável unidimensional.

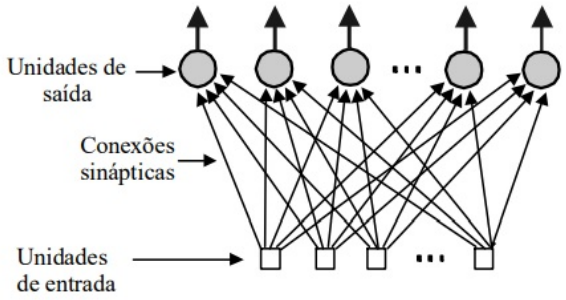

Figura 11 – Rede neural auto-organizável unidimensional (ZUBEN, 2004).

A seguir serão explicitados três processos subsequentes para formação dos mapas: processo competitivo, cooperativo e adaptativo.

4.5.1 Processo Competitivo

Neste processo, é calculada a similaridade entre cada neurônio da rede e o vetor de entrada. Conforme descrito em Souza (2012), este cálculo pode ser realizado através da distância euclidiana (equação 4.4) ou pelo método do cosseno (equação 4.5), da mesma forma como foi apresentado no cálculo da similaridade para sistemas RI.

$$D_{Euclidiana}(\vec{d_i}, p\vec{s}_j) = \sqrt{\sum_{k=1}^{n}(d_{i,k} - ps_{j,k})^2} \tag{4.4}$$

$$\cos(\vec{d_i}, p\vec{s}_j) = \frac{\sum_{k=1}^{n}(d_{i,k} * ps_{j,k})}{\sqrt{\sum_{k=1}^{n}(d_{i,k})^2 \times \sum_{k=1}^{n}(ps_{j,k})^2}} \tag{4.5}$$

Sendo que $d_{i,k}$ é o segmento k do vetor de entrada i e $ps_{j,k}$ é o segmento k do vetor de pesos sinápticos do neurônio j (SOUZA, 2012). A maior similaridade está associada ao maior valor do cosseno, calculado pela equação 4.4, ou pelo menor valor distância euclidiana, calculada pela equação 4.5. O neurônio com maior valor de similaridade é o mais apto a representar o padrão de entrada e, apenas este terá seu peso sináptico atualizado.

4.5.2 Processo Cooperativo

Esse processo baseia-se no mecanismo neurobiológico, no qual, quando um neurônio é ativado, o mesmo passa a influenciar os neurônios vizinhos. A vizinhança é dada por uma função gaussiana, calculada pela equação 4.6.

$$h_{j,i(d)} = exp(-\frac{d_{j,i}^2}{2\sigma^2}) \tag{4.6}$$

Onde $d_{j,i}$ é a distância topológica no mapa entre o neurônio vencedor i e o neurônio vizinho j e

σ é o desvio padrão da amplitude da vizinhança. EM geral o desvio padrão mostra-se relevante nos primeiros estágios do treinamento. Com o tempo de treinamento aumentando, o desvio padrão é controlado de acordo com a equação 4.7.

$$\sigma(n) = \sigma_0 exp(\frac{-n}{\tau_i}) \tag{4.7}$$

Sendo que n é intervalo de tempo discreto, σ_0 é uma constante definida pelo usuário que define o tamanho inicial da vizinhança e τ_i é uma constante de tempo que depende do número de iterações N, conforme equação 4.8.

$$\tau_i = \frac{N}{\log \sigma_0} \tag{4.8}$$

4.5.3 Processo Adaptativo

O processo adaptativo é a etapa onde ocorre o aprendizado da rede, através da atualização dos pesos sinápticos dos neurônios (SOUZA, 2012). O valor do peso sináptico $ps_j(n+1)$ do neurônio j, pertencente à vizinhança h_i do neurônio vencedor i, no instante imediatamente posterior ao atual é dado pela equação 4.9.

$$ps_j(n+1) = ps_j(n) + \eta(n) h_{j,i(d)}(n)(d_i - ps_j(n)) \tag{4.9}$$

Onde $d_i - ps_j(n)$ é o valor da taxa de erro, calculada pela subtração dos segmentos do vetor de entrada pelo vetor de pesos sinápticos do neurônio j e $\eta(n)$ é a taxa de aprendizado, calculada de acordo com a equação 4.10 (SOUZA, 2012).

$$\eta(n) = \eta_0 exp(-\frac{n}{\tau_2}) \tag{4.10}$$

Sendo que τ_2 é uma contante de tempo e η_0 é o valor inicial da taxa de aprendizagem.

4.5.4 Treinamento e Teste da Rede Neural

O treinamento da rede implementada por Souza (2012) é realizado em dois estágios: auto-organização e convergência. A auto-organização é a etapa na qual ocorrem a ordenação dos neurônios e o ajustes dos pesos sinápticos. A convergência é a etapa na qual ocorre um ajuste fino dos pesos, a fim de especializar ainda mais os neurônios, quando os mesmos padrões são apresentados à entrada.

Após o processo de auto-organização, dados de entrada com características semelhantes, produzirão reações semelhantes na rede. Dessa forma, comparando-se os efeitos da rede neural treinada, passa a ser possível agrupar os dados.

Na fase de teste, ocorre a classificação, na qual um conjunto de padrões de entrada é submetido à rede, que deve ser reconhecido e agrupado, a partir do treinamento realizado.

A metodologia empregada por Souza (2012) baseia-se em agrupar criptogramas através da rede neural com base na similaridade existente entre eles. O método é não-supervisionado, de modo que não há conhecimento em relação aos textos claros, os algoritmos criptográficos e a chave utilizada. A coleção de criptogramas é modelada no espaço vetorial, a matriz de similaridade é calculada pela medida do cosseno e o agrupamento é realizado por ligação simples. Finalmente, o agrupamento é avaliado pelas medidas de revogação e precisão.

A rede neural implementada em Souza (2012) conseguiu identificar com sucesso as cifras finalistas do concurso do AES (MARS, RC6, Rinjdael, Serpent e Twofish) de uma coleção de criptogramas. Além disso, o trabalho provou matematicamente que a metodologia é eficaz para um conjunto maior de cifras.

Outra contribuição importante do trabalho de Souza (2012) é a demonstração de que qualquer metodologia que baseia-se no conjunto léxico de uma linguagem para agrupar e classificar textos claros é capaz de agrupar e classificar criptogramas. Foi possível verificar que a combinação de parâmetros criptográficos geram um contexto linguístico, que pode ser entendido como uma assinatura no criptograma.

4.6 ALGORITMOS GENÉTICOS

Os algoritmos genéticos fazem parte de um grupo de modelos computacionais que inspiram-se no processo evolutivo das espécies, estudado pela biologia. As entradas do sistema são modeladas como cromossomos, que são submetidos a operações de que simulam a mutação e recombinação gênica.

O processamento de um algoritmo genético inicia-se com uma população aleatória de cromossomos. Cada cromossomo é avaliado e recebe um valor de probabilidade de reprodução, de modo que aqueles que representam uma melhor solução para o problema recebem valores mais altos de probabilidade do que aqueles que possuem uma solução pior.

4.6.1 Representação Cromossomial

A probabilidade de reprodução de cada cromossomo está associado a sua aptidão, que é calculada através de uma função objetivo. Essa função é construída a partir de parâmetros que podem ser conflitantes, ou seja, quando um aumenta o outro diminui. Dessa forma, o objetivo é encontrar um valor ótimo. A figura 12 ilustra o processo básico de um algoritmo genético.

O algoritmo genético é uma ferramenta computacional usada em problemas de busca de soluções em um espaço muito grande. O seu uso em busca de padrões criptográficos é justificado, visto que para um bloco com 128 *bits*, como no caso do AES, existem 2^{128} possibilidade de entrada.

Em seu trabalho, Oliveira e Xexéo (2013) define como a fase de pré-processamento, a etapa necessária para modelar os criptogramas em uma estrutura a ser utilizada como entrada

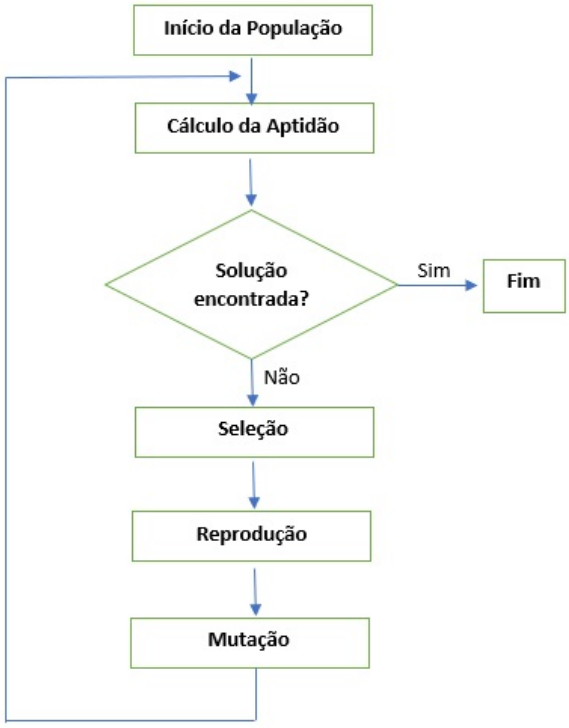

Figura 12 – Algoritmo Genético.

do algoritmo Genético. Essa fase equivale às etapas de modelagem dos criptogramas em um espaço vetorial e montagem da matriz de similaridade, calculada pelo cosseno, como descrito nas seções 4.2 e 4.3, respectivamente. Após essa fase, o algoritmo gera uma matriz binária aleatória que é a representação de um cromossomo da população inicial. Cada cromossomo é uma suposição de qual grupo cada criptograma deve pertencer. A figura 13 ilustra um exemplo de representação cromossomial.

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

Figura 13 – Representação Cromossomial. Adaptado de Oliveira e Xexéo (2013).

Os criptogramas estão representados nas linhas da matriz, enquanto que as colunas representam grupos. Para cada elemento a_{ij} iguais a um, tem-se que o criptograma i pertence ao grupo j, caso contrário a_{ij} será igual a zero. Vale ressaltar que cada criptograma só pode pertencer a apenas um único grupo.

4.6.2 Funcionamento do Algoritmo Genético

A avaliação de cada cromossomo é calculada pela função *Calinski-Harabazs* (CALINSKI; HARABASZ, 1974), a fim de determinar a qualidade de cada um. A função está descrita pela equação 4.11.

$$CH = \frac{\sum_{k=1}^{K} n_K \|z_k - z\|^2 \times (n-k)}{\sum_{k=1}^{K} \sum_{i=1}^{n_K} \|x_i - z_k\|^2 \times (n-1)} \tag{4.11}$$

Onde K é a quantidade de grupos de uma coleção, n_k é quantidade de criptogramas do grupo k, z_k é o centro geométrico do grupo k no espaço vetorial, z é centro geométrico de todos os criptogramas e x_i é o *i-ésimo* criptograma de um conjunto.

O funcionamento do algoritmo genético implementado em Oliveira e Xexéo (2013) inicia-se com a geração de 200 cromossomos aleatórios, de acordo com o modelo da figura 13. Os passos de cada iteração do algoritmo estão descritos a seguir:

a) Determinação do melhor cromossomo da população pela função *Calinski-Harabazs*;

b) Formação da segunda população da geração n, denominada "população sorteada", composta pelo cromossomo eleito no item b e mais 199 cromossomos sorteados da primeira população. No sorteio, alguns cromossomos podem ser repetidos ou eliminados;

c) Criação da primeira população da geração $n+1$, através de operações de *crossover* e mutação entre os cromossomos;

d) Descarte e substituição do pior cromossomo da primeira população da geração $n+1$ pelo melhor cromossomo da primeira população da geração n; e

e) Repetir o processo até alcançar o total de iterações.

A etapa relativa ao item b corresponde a um método denominado "elitismo". A ação de inserir o melhor cromossomo na população seguinte evita a perda de informações importantes presentes em indivíduos de alta avaliação e que poderiam ser perdidas nos processos de seleção e cruzamento.

O agrupamento de criptogramas é realizado em uma fase de treinamento. Nessa fase, uma coleção de criptogramas conhecidos é aplicada como entrada do algoritmo. Os criptogramas agrupados formam um dicionário de blocos. Já na fase de teste ocorre a classificação, através da interseção de um criptograma desconhecido e o dicionários de blocos binários. Se houver interseção, o criptograma desconhecido será classificado, caso contrário, não haverá classificação Oliveira e Xexéo (2013). A figura 14 ilustra o processo.

O algoritmo genético implementado no trabalho de Oliveira e Xexéo (2013) obteve sucesso na separação e identificação das cinco cifras finalistas do concurso do AES. Como no trabalho de Souza (2012), a identificação das cifras demonstra a existência de assinaturas decorrentes das operações realizadas pelos algoritmos criptográficos ou provocadas pelas chaves.

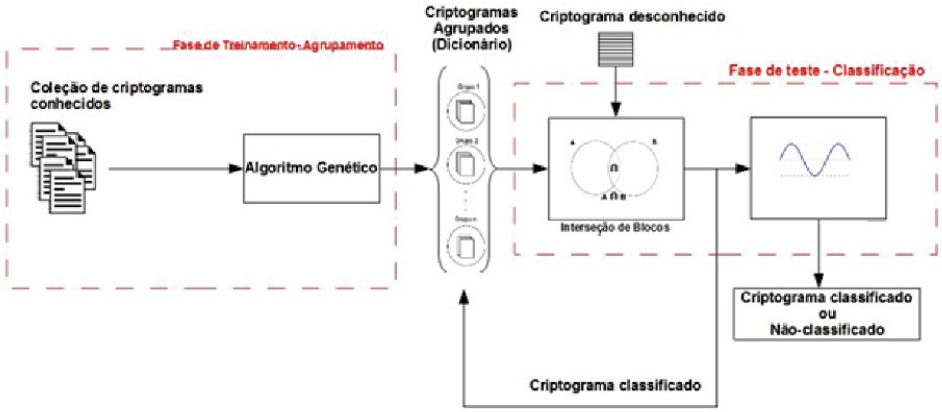

Figura 14 – Agrupamento e Classificação do Algoritmo GenéticoOliveira e Xexéo (2013).

5 CONCLUSÕES

Esse trabalho apresenta um conjunto de trabalhos recentes referentes a técnicas não convencionais de criptoanálise, que em paralelo com as técnicas convencionais, servem como ferramentas criptoanalíticas. O principal objetivo é contribuir com um documento que reúne essas técnicas e que funciona como um ponto de partida para pesquisas na área de criptoanálise. Essa contribuição está alinhada com a Doutrina de Tecnologia da Informação, que possui como um dos seus propósitos "promover a capacitação de recursos humanos na MB, para o desenvolvimento de competência científico-tecnológica em TI e criptologia, envolvendo tanto as atividades de desenvolvimento de códigos quanto as atividades de criptoanálise"(BRASIL, 2007).

O presente trabalho compromete-se também com o propósito de fortalecimento da mentalidade de segurança digital no âmbito militar-naval. Em especial, contribuir com o entendimento de que um sistema criptográfico nunca será considerado absolutamente seguro e só poderá ser considerado robusto se submetido a testes criptoanalíticos. Dessa forma, ao se implantar um novo sistema criptográfico, deve-se avaliar a quais tipos de ataques, o sistema foi submetido e a quais o mesmo foi reprovado. Tal necessidade fica evidente pela própria Doutrina de TI da Marinha do Brasil, ao apresentar uma de suas atividades de TI: "execução de atividades de criptoanálise operacional e certificacional"(BRASIL, 2007).

A criptoanálise certificacional compreende os serviços relativos à verificação da robustez das cifras utilizadas, bem como serviços de certificação e homologação dos sistemas criptográficos. Novamente, as técnicas apresentadas, além dos métodos convencionais, servem como subsídios para submeter os recursos criptográficos da Marinha a testes rigorosos de criptoanálise, a fim de garantir a robustez de suas cifras.

Entende-se por criptoanálise operacional, as tarefas de inteligência, interceptação de mensagens criptografadas e decifração. Ainda que a extração completa de uma mensagem em texto claro a partir de um criptograma interceptado seja muito improvável, as técnicas apresentadas de identificação, classificação e agrupamento de cifras podem fornecer subsídios importantes para serviços de inteligência.

Através da análise de tráfego, o serviço de inteligência da Marinha pode capturar informações sobre padrões de tráfego de dados, bem como frequência e tamanho de mensagens, ainda que criptografadas. Entretanto, com posse das metodologias apresentadas neste trabalho, em especial as técnicas baseadas em recuperação das informações, pode-se obter informações ainda mais valiosas, tais como identificação de cifras, ou a identificação das mudanças de algoritmos utilizados na cifração (SOUZA, 2012).

Os métodos baseados em recuperação da informação, como a rede neural desenvolvida por Souza (2012) e o algoritmo genético desenvolvido por Oliveira e Xexéo (2013) foram testados com criptogramas gerados pelos algoritmos finalistas do concurso do AES. Testes com

outras coleções de criptogramas, gerados por recursos criptográficos utilizados pela Marinha, como o RSA, poderiam ser a base de estudos futuros. Além disso, o trabalho de Souza (2012) provou que qualquer técnica de agrupamento e classificação de textos pode ser utilizada para agrupar e classificar criptogramas. Dessa forma, ainda que estudos complementares sejam necessários, julga-se viável e vantajosa a adoção de técnicas baseadas em RI como ferramentas de criptoanálise certificacional e criptoanálise operacional da Marinha do Brasil.

REFERÊNCIAS

ALVARENGA, L. G. D. **Criptografia Clássica e Moderna**. [S.l.: s.n.], 2010. 366 p.

BIHAM, E.; SHAMIR, A. **Differential cryptanalysis of DES-like cryptosystems**. Journal of CRYPTOLOGY, Springer, v. 4, n. 1, p. 3–72, 1991.

BRASIL. Estado Maior da Armada. **Doutrina de Tecnologia da Informação da Marinha**. 2007.

CALINSKI, T.; HARABASZ, J. **A dendrite method for cluster analysis.** Communications in Statistics, 3(1), 1974, 1-27, 1974, 1974.

CARVALHO, C. A. B. d. **O uso de técnicas de recuperação de informações em criptoanálise**. 79 f. Tese (Mestrado em Sistemas e Computação) — Instituto Militar de Engenharia, Rio de Janeiro., Rio de Janeiro, 2006.

CHANDRA, G.; THE, T. **Classication Of Modern Ciphers**. 2002.

CRYPTO. **Criptografia de dados e gerenciamento de chaves**. 2020. Disponível em: <https://cryptoid.com.br/e-val-tecnologia/criptografia-de-dados-e-gerenciamento-de-chaves-ouca/>.

DILEEP, A. D.; SEKHAR, C. C. **Identification of block ciphers using support vector machines.** In International Joint Conference on Neural Networks (IJCNN 2006), Vancouver, BC, Canada, July 2006, pp. 2696-2701., 2006.

FRAKES, W. B. **Introduction to information storage and retrieval system.** In: FRAKES, William B, YATES, Ricardo B. Information retrieval: data structures and algorithms. Upper Saddle River: Prentice Hall, 1992. p. 1-12. Trabalho Individual, UFRGS, 1997.

FUNG, B. C. M. **Hierarchical document clustering using frequent itemsets.** Proceedings of the SIAM International Conference on Data Mining, (SDMb 2003), May 2003, San Francisco, CA., 2003.

HARMAN, D. **Ranking algorithms.** In: FRAKES, William B, YATES, Ricardo B. Information retrieval: data structures and algorithms. Upper Saddle River: Prentice Hall, 1992. p. 363-392., 1992.

KNUDSEN, L. R.; MEIER, W. **Correlations in RC6 with a Reduced Number of Rounds**. In: GOOS, G. et al. (Ed.). Fast Software Encryption. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001. p. 94–108.

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

MAHESHWARI, P. **Classification of ciphers**. M. Tech Thesis. Department of Computer Science and Engineering, Indian Institute of Technology. Kanpur, 2001.

MATSUI, M. **Linear cryptanalysis method for DES cipher**. [S.l.: s.n.].

NIST. NATIONAL INSTITUTE OF STANDARD AND TECNOLOGY. **Federal Information Processing Standard, publication 46-3 (FIPS 46-3): Data Encryption Standard (DES).** . 1999.

NIST. NATIONAL INSTITUTE OF STANDARD AND TECNOLOGY. **Federal Information Processing Standard, publication 197 (FIPS 197): Announcing the advanced encryption standard (AES)**. 2001.

NIST. NATIONAL INSTITUTE OF STANDARD AND TECNOLOGY. **A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications.** NIST Special Publication 800-22. Revision 1. 2008.

OLIVEIRA, G. A. de; XEXéO, J. A. M. **A Aplicação de Algoritmos Genéticos no Reconhecimento de Padrões Criptográficos**. Seção de Engenharia de Sistemas – Instituto Militar de Engenharia, 2013.

RASMUSSEN, E. **Clustering algorithms**. In: FRAKES, William B, YATES, Ricardo B. Information retrieval: data structures and algorithms. Upper Saddle River: Prentice Hall, 1992. p. 419-442., 1992.

RIBEIRO, C.; ROIHA, L. **Estudo Comparativo dos Modos de Operação de Confidencialidade: um Overview para Iniciantes**. *Revista Ciência e Tecnologia*, v. 8, n. 13, 2010. ISSN 2236-6733. Disponível em: <http://www.revista.unisal.br/sj/index.php/123/article/view/70>.

RUKHIN, A. E. A. **A statistical test suite for the validation of cryptographic random number generators**. 1999.

SOUZA, W. A. R. **Identificação de padrões em criptogramas usando técnicas de classificação de textos**. 252 f. Tese (Mestrado em Sistemas e Computação) — Instituto Militar de Engenharia, Rio de Janeiro, 2007.

SOUZA, W. A. R. **Identificação de contextos linguísticos em linguagens desconhecidas geradas por cifras de bloco.** 105 f. Tese (Doutorado Engenharia de Sistemas e Computação) — COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2011.

SOUZA, W. A. R. **Identificação de Cifras de Blocos por meio de Rede neural Artificial**. 48 f. Tese (Especialista em Criptografia) — Universidade Federal Fluminense, Niterói, 2012.

STEVENSON, D. **1914-1918: A história da Primeira Guerra Mundial**. São Paulo: Novo Século, 2016. 752 p.

TERADA, R.; UEDA, E. T. **A new version of the RC6 algorithm, stronger against χ^2 cryptanalysis**. Proc. 7th Australasian Information Security Conference (AISC 2009), Wellington, New Zealand, 2009.

VAUDENAY, S. **An Experiment on DES Statistical Cryptanalysis**. In: Proceedings of the 3rd ACM Conference on Computer and Communications Security. New York, NY, USA: Association for Computing Machinery, 1996. (CCS '96), p. 139–147. ISBN 0897918290. Disponível em: <https://doi.org/10.1145/238168.238206>.

WIVES, L. K. **Técnicas de descoberta de conhecimento em textos aplicadas à inteligência competitiva**. Trabalho Individual, UFRGS, 1997.

YATES, R. B.; NETO., R. **Modern information retrieval**. New York: addison Wesley, 1999.

ZUBEN, F. J. V. **Máquinas de vetores-suporte**. *DCA/FEEC/Unicamp*, 2003. Disponível em: <ftp://ftp.dca.fee.unicamp.br/pub/docs/vonzuben/ia353_1s13/topico8_1s2013.pdf>.

ZUBEN, F. J. V. **Introdução à Computação Natural**. *DCA/FEEC/Unicamp*, 2004. Disponível em: <ftp://ftp.dca.fee.unicamp.br/pub/docs/vonzuben/ia006_03/topico1_03.pdf>.