

**UNIVERSIDADE ESTADUAL PAULISTA**  
**“Júlio de Mesquita Filho”**  
Pós-Graduação em Ciência da Computação

Luis Alexandre da Silva

**Aprendizado Não-Supervisionado de Características para  
Detecção de Conteúdo Malicioso**

São José do Rio Preto  
2016

Luis Alexandre da Silva

# **Aprendizado Não-Supervisionado de Características para Detecção de Conteúdo Malicioso**

Dissertação de Mestrado elaborada junto ao Programa de Pós-Graduação em Ciência da Computação - Área de Concentração em Computação Aplicada, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Prof. Dr. João Paulo Papa

Orientador

Prof. Dr. Kelton Augusto Pontara da Costa

Co-orientador

São José do Rio Preto

2016

Silva, Luis Alexandre da.

Aprendizado não-supervisionado de características para detecção de conteúdo malicioso / Luis Alexandre da Silva. -- São José do Rio Preto, 2016

57 f.: il.

Orientador: João Paulo Papa

Dissertação (mestrado) – Universidade Estadual Paulista “Júlio de Mesquita Filho”, Instituto de Biociências, Letras e Ciências Exatas

1. Computação. 2. Biometria. 3. Reconhecimento de padrões. 4. Processamento de imagens - Técnicas digitais. 5. Redes neurais - (Computação) 6. Aprendizado do computador. I. Universidade Estadual Paulista "Júlio de Mesquita Filho". Instituto de Biociências, Letras e Ciências Exatas. II. Título.

CDU - 578.087

Ficha catalográfica elaborada pela Biblioteca do IBILCE  
UNESP - Câmpus de São José do Rio Preto

Luis Alexandre da Silva

# **Aprendizado Não-Supervisionado de Características para Detecção de Conteúdo Malicioso**

Dissertação de Mestrado elaborada junto ao Programa de Pós-Graduação em Ciência da Computação - Área de Concentração em Computação Aplicada, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Comissão Examinadora

Prof. Dr. João Paulo Papa

Faculdade de Ciências - Departamento de Computação

Universidade Estadual Paulista - Bauru

(Presidente)

Profa. Dra. Roberta Spolon

Faculdade de Ciências - Departamento de Computação

Universidade Estadual Paulista - Bauru

Prof. Dr. Tiago Agostinho de Almeida

Departamento de Computação

Universidade Federal de São Carlos - Sorocaba

São José do Rio Preto

25 de agosto de 2016

*À minha família, em especial aos meus pais João e Júlia, a minha esposa Ana Paula e minha filha Ana Laura, por todo amor, apoio, confiança e incentivo em todos os momentos.*

## **AGRADECIMENTOS**

Meus agradecimentos a todos os familiares, amigos, professores e funcionários ligados ao IBILCE-UNESP, que direta ou indiretamente contribuíram para a realização deste trabalho. Em especial, dedico meus agradecimentos:

- A Deus, por ter me dado força e saúde para chegar até aqui;
- Aos meus pais João e Júlia pelo carinho, apoio e incentivo;
- A minha esposa Ana Paula e filha Ana Laura pelo amor, apoio, confiança e incentivo em todos os momentos;
- Ao Prof. Dr. João Paulo Papa, por todo ensinamento, incentivo, confiança e orientação;
- Ao Prof. Dr. Kelton Augusto Pontara da Costa pela co-orientação, todo o ensinamento e pela amizade;
- Aos Professores Dra. Roberta Spolon e Dr. Tiago Agostinho de Almeida, pelo acompanhamento nas bancas examinadoras, sugestões e incentivos;
- Aos meus amigos e colegas que de forma direta ou indiretamente me ajudaram, em especial ao Gustavo César Bruschi.

*“O sol é para todos,”  
mas a sombra é para quem.*

*chega primeiro.*

***Geremias Ludu***

*“Um pouco de ciência nos afasta de Deus. Muito, nos aproxima.”*

***Louis Pasteur (1822-1895)***



## RESUMO

O aprendizado de características tem sido um dos grandes desafios das técnicas baseadas em Redes Neurais Artificiais (RNAs), principalmente quando se trata de um grande número de amostras e características que as definem. Uma técnica ainda pouco explorada nesse campo diz respeito as baseadas em RNAs derivada das Máquinas de Boltzmann Restritas, do inglês *Restricted Boltzmann Machines* (RBM), principalmente na área de segurança de redes de computadores. A proposta deste trabalho visa explorar essas técnicas no campo de aprendizado não-supervisionado de características para detecção de conteúdo malicioso, especificamente na área de segurança de redes de computadores. Experimentos foram conduzidos usando técnicas baseadas em RBMs para o aprendizado não-supervisionado de características visando a detecção de conteúdo malicioso utilizando meta-heurísticas baseadas em algoritmos de otimização, voltado à detecção de *spam* em mensagens eletrônicas. Nos resultados alcançados por meio dos experimentos, observou-se, que com uma quantidade menor de características, podem ser obtidos resultados similares de acurácia quando comparados com as bases originais, com um menor tempo relacionado ao processo de treinamento, evidenciando que técnicas de aprendizado baseadas em RBMs são adequadas para o aprendizado de características no contexto deste trabalho.

**Palavras-chave:** Aprendizado de Características. Anomalias. Redes de Computadores. Redes Neurais Artificiais. Máquinas de Boltzmann Restritas.

## ABSTRACT

The features learning has been one of the main challenges of techniques based on Artificial Neural Networks (ANN), especially when it comes to a large number of samples and features that define them. Restricted Boltzmann Machines (RBM) is a technique based on ANN, even little explored especially in security in computer networks. This study aims to explore these techniques in unsupervised features learning in order to detect malicious content, specifically in the security area in computer networks. Experiments were conducted using techniques based on RBMs for unsupervised features learning, which was aimed to identify malicious content, using meta-heuristics based on optimization algorithms, which was designed to detect spam in email messages. The experiment results demonstrated that fewer features can get similar results as the accuracy of the original bases with a lower training time, it was concluded that learning techniques based on RBMs are suitable for features learning in the context of this work.

**Keywords:** Feature Learning. Anomalies. Computer Networks. Artificial Neural Networks. Restricted Boltzmann Machines.

## LISTA DE FIGURAS

Figura 1	Arquitetura de uma RBM. . . . .	26
Figura 2	Amostragem de Gibbs para estimar o modelo de reconstrução dos dados. . .	29
Figura 3	Ilustração do método baseado em divergência contrastiva. . . . .	30

## LISTA DE TABELAS

Tabela 1	Algoritmo Classificador Supervisionado baseado em Floresta de Caminhos Ótimos usando grafo completo. . . . .	35
Tabela 2	Configuração dos parâmetros meta-heurísticos de otimização baseados em técnicas de Busca Harmônica. . . . .	38
Tabela 3	Resultados das configurações CONF1 e CONF2 do experimento EXP1 da SPAMBASE. . . . .	43
Tabela 4	Resultados das configurações CONF1 e CONF2 do experimento EXP1 da LINGSPAM. . . . .	44
Tabela 5	Resultados das configurações CONF1 e CONF2 do experimento EXP2 da LINGSPAM. . . . .	45
Tabela 6	Resultados das configurações CONF1 e CONF2 do experimento EXP1 da CSDMC. . . . .	46
Tabela 7	Resultados das configurações CONF1 e CONF2 do experimento EXP2 da CSDMC. . . . .	46
Tabela 8	Resultados consolidados dos experimentos EXP1 e EXP2. . . . .	47
Tabela 9	Resultados das medições de tempos médios de treinamento e teste dos experimentos EXP1. . . . .	47
Tabela 10	Resultados das medições de tempos médios de treinamento e teste dos experimentos EXP2. . . . .	48

## LISTA DE ABREVIATURAS E SIGLAS

BRBM	Bernoulli Restricted Boltzmann Machines.
CD	Contrastive Divergence.
DBN	Deep Belief Networks.
DBN-C	Deep Belief Network on Channel epochs.
DBN-T	Deep Belief Network on Time Samples.
DBN-W	Deep Belief Network on Windowed samples.
DF	Document Frequency Thresholding.
FPCD	Fast Persistent Contrastive Divergence.
HMRC	Harmony Memory Considering Rate.
HS	Harmony Search.
IDS	Intrusion Detection Systems.
IFT	Image Foresting Transform.
IG	Information Gain.
IP	Internet Protocol.
MI	Mutual Information.
MST	Minimum Spanning Tree.
NLP	Natural Language Processing.
OPF	Optimum-Path Forest.
PAR	Pitch Adjusting Rate.
PCA	Principal Component Analysis.
PCD	Persistent Contrastive Divergence.
RBM	Restricted Boltzmann Machines.
RNA	Redes Neurais Artificiais.
TS	Term Strength.
URL	Uniform Resource Locator.

## SUMÁRIO

<b>1</b>	<b>INTRODUÇÃO</b>	<b>14</b>
1.1	TRABALHOS RELACIONADOS	15
1.2	OBJETIVOS	17
<b>1.2.1</b>	<b>Objetivos Gerais</b>	<b>17</b>
<b>1.2.2</b>	<b>Objetivos Específicos</b>	<b>17</b>
1.3	MOTIVAÇÃO	18
1.4	JUSTIFICATIVAS	18
1.5	ESTRUTURA DO TRABALHO	18
<b>2</b>	<b>FUNDAMENTAÇÃO TEÓRICA</b>	<b>20</b>
2.1	SEGURANÇA DE REDES DE COMPUTADORES	20
<b>2.1.1</b>	<b>Detecção de Spam</b>	<b>21</b>
2.2	REDUÇÃO DE DIMENSIONALIDADE	23
<b>2.2.1</b>	<b>Seleção de Características</b>	<b>24</b>
<b>2.2.1.1</b>	<b><i>Distribuição <math>\chi^2</math></i></b>	<b>25</b>
2.3	MÁQUINAS DE BOLTZMANN RESTRITAS	26
<b>2.3.1</b>	<b>Divergência Contrastiva</b>	<b>29</b>
<b>2.3.2</b>	<b>Divergência Contrastiva Persistente</b>	<b>31</b>
<b>2.3.3</b>	<b>Divergência Contrastiva Persistente Rápida</b>	<b>32</b>
2.4	FLORESTA DE CAMINHOS ÓTIMOS	33
<b>2.4.1</b>	<b>Grafo Completo</b>	<b>34</b>
<b>2.4.2</b>	<b>Algoritmo de Treinamento</b>	<b>35</b>
<b>2.4.3</b>	<b>Método de Classificação</b>	<b>36</b>

<b>3</b>	<b>MATERIAIS E MÉTODOS</b>	<b>37</b>
3.1	BASES DE DADOS	39
3.2	PROTOCOLO DOS EXPERIMENTOS: EXP1	40
3.3	PROTOCOLO DOS EXPERIMENTOS: EXP2	41
<b>4</b>	<b>RESULTADOS EXPERIMENTAIS</b>	<b>42</b>
4.1	RESULTADOS EXPERIMENTOS: SPAMBASE	43
4.2	RESULTADOS EXPERIMENTOS: LINGSPAM	44
4.3	RESULTADOS EXPERIMENTOS: CSDMC	45
4.4	CONSIDERAÇÕES FINAIS	46
<b>5</b>	<b>CONCLUSÕES</b>	<b>49</b>
5.1	PRINCIPAIS CONTRIBUIÇÕES	50
5.2	TRABALHOS FUTUROS	51
<b>6</b>	<b>PUBLICAÇÕES REALIZADAS</b>	<b>52</b>
	<b>REFERÊNCIAS</b>	<b>54</b>

## 1 INTRODUÇÃO

Nas últimas décadas, tem-se observado um grande aumento do volume de informações armazenadas e disponibilizadas, o que tem causado dificuldades relacionadas à classificação e identificação dessas informações diante desse volume de dados. Nesse contexto, podem ser destacados diversos conteúdos maliciosos em redes de computadores, que têm aumentado consideravelmente devido ao grande tráfego das redes, levantando uma preocupação direcionada à criação de técnicas para identificar esses padrões de tráfego incomuns. Nessa linha, técnicas como Redes Neurais Artificiais (RNAs)<sup>1</sup> têm sido utilizadas para realizar tarefas de aprendizado de máquina. As arquiteturas de RNAs foram projetadas utilizando as ideias da biologia com o intuito de simular o funcionamento do cérebro (LOPES; RIBEIRO; GONCALVES, 2012). As técnicas baseadas em arquiteturas rasas apresentam, geralmente, apenas uma camada oculta, enquanto que as arquiteturas profundas possuem várias camadas ocultas (multi-camadas), podendo ser melhores alternativas para conjuntos de dados que apresentam uma grande quantidade de características (HINTON; SALAKHUTDINOV, 2006; LOPES; RIBEIRO; GONCALVES, 2012). Tais arquiteturas entraram em evidência devido a novos resultados empíricos e teóricos nos últimos tempos, elevando o interesse dos pesquisadores no domínio das redes neurais (LOPES; RIBEIRO; GONCALVES, 2012). Muitos trabalhos têm utilizado estes tipos de redes neurais, como as Máquinas de *Boltzmann* Restritas, do inglês *Restricted Boltzmann Machines* (RBM) (ACKLEY; HINTON; SEJNOWSKI, 1988; HINTON, 2002; LAROCHELLE; BENGIO, 2008; HINTON, 2012; LAROCHELLE et al., 2012; FISCHER; IGEL, 2014; PAPA et al., 2015b), tendo aplicabilidades de sucesso em vários domínios, incluindo classificação, regressão, redução de dimensionalidade, segmentação de objetos, recuperação de informação, robótica, processamento de linguagem natural e filtragem colaborativa, entre outros (LOPES; RIBEIRO; GONCALVES, 2012).

Arquiteturas RBMs treinadas de forma não-supervisionada também têm sido propostas como um método automático para o aprendizado de características discriminativas. Dessa forma, estes modelos de aprendizagem podem aprender características mais discriminativas para um determinado problema (HE et al., 2013), e eventualmente melhorar o custo computacional e o tempo necessário para realizar o processo de treinamento (aprendizado).

---

<sup>1</sup>Modelo computacional inspirado no funcionamento do cérebro humano, sendo geralmente representado como um sistema de neurônios interconectados que podem processar valores de entradas (PAN et al., 2014).



A área de segurança de redes tem como objetivo preocupar-se com diversos métodos malignos de intrusões que possam ingressar na rede e provocar danos. Nas últimas décadas, houve um aumento na quantidade de diferentes ataques em redes de computadores, levando a uma preocupação com a identificação destes, principalmente com o crescimento exponencial das informações disponíveis e o aumento do tráfego nas redes, tornando cada vez mais importante encontrar formas de classificá-las, identificá-las e acessá-las. Entretanto, é importante encontrar as características que melhor definem tais informações, sendo esse um dos grandes desafios relacionadas ao emprego das técnicas baseadas em aprendizado.

Na área de segurança de redes, tais dificuldades podem ser observadas quando são abordadas as detecções de conteúdos maliciosos, devido ao volume e a complexidade do tráfego de rede. Dentre as anomalias encontradas em redes de computadores, pode-se destacar a detecção de *spam* no contexto de mensagens eletrônicas, no caso e-mails, que vem sendo bem explorada nas técnicas de aprendizado de máquina, a qual será tratada nos experimentos deste trabalho.

De maneira geral, este trabalho tem como proposta empregar técnicas de arquiteturas baseadas em RBM, treinadas de forma não-supervisionada, como um método para aprender características mais discriminativas, especificamente na área de segurança de redes de computadores.

## 1.1 TRABALHOS RELACIONADOS

No contexto deste estudo é importante destacar alguns trabalhos que seguem na mesma linha de pesquisa. Voltado à detecção de anomalias em redes de computadores, Fiore et al. (2013) empregaram RBM como um instrumento para inspecionar o tráfego de rede e discriminar entre comportamentos anômalos e normais para detectar atividades indesejadas ou suspeitas, não com o objetivo de criar um sistema de detecção de intrusão e reação em tempo real, mas no sentido de uma descrição melhor e mais adequada de tráfego de rede. Para isso, exploraram a eficácia de uma abordagem semi-supervisionada utilizando apenas a classe normal para o aprendizado, já que essa compreende a maior parte do tráfego. Os experimentos foram realizados com base em dois conjuntos de dados de tráfego de rede: KDD 1999<sup>2</sup> e amostras coletadas do tráfego de uma rede de computadores controlada, representando um tráfego real. Os experimentos foram conduzidos por meio de um cruzamento de treinamento e teste entre os dois conjuntos de dados, sendo que as experiências confirmaram que o uso da RBM apresentou bons resultados para capturar os aspectos inerentes à classe de tráfego normal.

Hajinoroozi et al. (2015) propõem um método de aprendizado de características e redução de dimensionalidade com o uso de Redes Neurais de Crenças Profundas, do inglês *Deep Belief*

---

<sup>2</sup><http://www.sigkdd.org/kddcup/index.php?section=1999>

*Networks* (DBN) (HINTON; OSINDERO; TEH, 2006; ZHOU; CHEN; WANG, 2010; SARIKAYA; HINTON; DEORAS, 2014; PAPA; SCHEIRER; COX, 2015), derivadas das RBMs, utilizando um conjunto de dados baseado em eletroencefalografia relativo ao prognóstico do estado mental de percepção de motoristas. O trabalho investiga como DBN em modo não-supervisionado pode aprender melhores características visando aumentar a performance de classificação dos estados de percepção. Para isso, os autores usaram três métodos baseados em DBN: *Deep Belief Network on Time Samples* (DBN-T), *Deep Belief Network on Windowed samples* (DBN-W) e *Deep Belief Network on Channel epochs* (DBN-C). Os resultados dos experimentos evidenciaram que o uso de DBN no aprendizado de características obteve melhores resultados do que os alcançados pelos métodos tradicionais. Dentre os métodos utilizados, DBN-C alcançou os melhores resultados.

Um dos maiores problemas relacionados ao Processamento de Linguagem Natural, do inglês *Natural Language Processing* (NLP), bem como ao aprendizado de características baseado em palavras, é a questão da dimensionalidade do vocabulário. Neste sentido, Dahl, Adams e Larochelle (2012) demonstraram uma abordagem utilizando treinamento baseado em RBMs sobre um vocabulário extenso de palavras usando o aprendizado de características para melhorar a performance sobre a tarefa de classificação. Neste trabalho, os autores descrevem uma regra de aprendizagem baseada em RBM com uma complexidade computacional independente do número de unidades visíveis, substituindo a função de transição baseada em amostragem de Gibbs (*Gibbs Sampling*) sobre as unidades visíveis por uma implementação baseada no algoritmo *Metropolis-Hastings Transitions*. Ao treinar uma RBM desta maneira, com centenas de milhões de palavras, são aprendidas representações da captura de propriedades sintáticas e semânticas significativas das mesmas. As representações das palavras aprendidas proporcionam benefícios na tarefa de organizar elementos individuais competitivos quando comparados à outros métodos de indução de representações de palavras. Como resultado, apresentam que as características aprendidas podem produzir ganhos de desempenho no processo de aprendizado.

Voltado à detecção de anomalias em redes de computadores, podemos destacar a detecção de mensagens de *spam* em e-mails através do uso de vocabulários (uso de palavras). Silva et al. (2015) propõem o uso da RBM como um meio de identificar mensagens de *spam* em mensagens de e-mail utilizando o aprendizado não-supervisionado de características. Neste trabalho, foram utilizados dois conjuntos de dados bem conhecidos relacionados a mensagens de e-mails: SPAMBASE e LINGSPAM. O conjunto de dados LINGSPAM foi pré-processado com o mesmo dicionário de palavras (*tokens*) da SPAMBASE; em seguida, as palavras foram binarizadas, indicando 1 (um) quando há a presença da palavra, ou 0 (zero) quando não há a presença. Neste caso, a RBM foi utilizada para realizar o aprendizado não-supervisionado das

características dos conjunto de dados. Para isso, foram utilizadas três técnicas baseadas em Busca Harmônica, do inglês *Harmony Search* (HS) (GEEM; KIM; LOGANATHAN, 2001), para refinamento dos parâmetros da RBM. Os experimentos foram conduzidos dividindo as amostras dos conjuntos. Já na camada de treinamento e classificação, foi utilizado o classificador baseado em Floresta de Caminhos Ótimos, do inglês *Optimum-Path Forest* (OPF) (PAPA et al., 2012), que após ser aplicado teve os resultados de acurácia obtidos sob os dados originais e os processados utilizando RBM comparados, mostrando que há um caminho promissor com o uso das RBMs no aprendizado de características. Outro relevante trabalho neste contexto foi descrito pelos pesquisadores Almeida e Yamakami (2012), que apresentaram uma nova abordagem baseada no princípio da descrição de comprimento mínimo. Os autores compararam a proposta com Máquina de Vetores de Suporte Linear e sete outros modelos distintos do classificador *Naïve Bayes*. É importante ressaltar que o filtro de *spam* proposto obteve o melhor desempenho médio para todos os conjuntos de dados analisados, alcançando uma taxa de precisão superior a 95% para todos os conjuntos de dados de e-mail.

## 1.2 OBJETIVOS

### 1.2.1 Objetivos Gerais

Este trabalho tem como objetivos gerais o emprego de técnicas de aprendizado baseadas em RBM visando aprender características de maneira não-supervisionada, bem como buscar resultados mais eficazes na detecção de conteúdo malicioso na área de segurança de redes de computadores, pela aplicação destes métodos.

### 1.2.2 Objetivos Específicos

Nesta proposta, destacam-se os seguintes objetivos específicos:

- Contribuir com a literatura da área de aprendizado de características usando RNA, especificamente com o uso das RBMs;
- Diminuir a dimensionalidade das características a serem computadas e, conseqüentemente, o custo computacional relacionado ao treinamento e classificação das amostras;
- Encontrar características mais discriminativas que definem as classes;
- Verificar a viabilidade da aplicação de técnicas de aprendizado de características baseadas em RBM no contexto do aprendizado não-supervisionado para à detecção de conteúdo malicioso em redes de computadores.

### 1.3 MOTIVAÇÃO

Com o crescimento exponencial das informações disponíveis no mundo, torna-se cada vez mais importante encontrar formas de classificá-las, identificá-las e acessá-las de maneiras mais rápidas e eficientes. Isso é perceptível nas redes de computadores devido ao grande número de dados que trafegam, pois existe uma grande dificuldade em separar conteúdos normais em relação aos conteúdos maliciosos, os quais podem ser uma ameaça à confiabilidade e segurança dos dados. Essas anomalias podem se manifestar de diversas formas, tais como: *spams*, *malwares* e invasões, dentre outros. Entretanto, é importante encontrar os atributos que melhor definem essas informações com o intuito de facilitar a identificação desses conteúdos maliciosos em relação aos conteúdos normais, sendo esse um dos grandes desafios relacionados ao emprego das técnicas baseadas em aprendizado exploradas neste trabalho.

### 1.4 JUSTIFICATIVAS

Uma das grandes dificuldades das técnicas de aprendizado é que, diante de uma grande quantidade de características, o processo de treinamento tem um alto custo computacional, sendo em alguns casos inviável realizar esta etapa em tempo hábil. Em contrapartida, o uso de poucos atributos pode tornar este processo ineficaz, prejudicando a etapa de classificação. Ou seja, torna-se cada vez mais importante a necessidade de aprender as informações mais discriminativas para o processo de aprendizado. Com o processo de aprendizado mais relevante, a etapa de treinamento pode ser realizada com um conjunto menor de atributos e ainda obter bons resultados.

Diante das dificuldades apresentadas no processo de treinamento, torna-se cada vez mais importante buscar formas de detectar características discriminativas, que melhor definam as amostras a serem trabalhadas. Um bom trabalho realizado nessa linha pode representar uma maior qualidade nas etapas de treinamento e, conseqüentemente, na classificação de novas amostras.

### 1.5 ESTRUTURA DO TRABALHO

O restante deste trabalho é organizado como segue. No Capítulo 2, são apresentados os conceitos teóricos importantes no desenvolvimento deste trabalho. Nos Capítulos 3 e 4 são descritos os materiais e métodos utilizados na condução dos experimentos realizados, bem como seus resultados e suas considerações, respectivamente. Já no Capítulo 5, são apresentadas as

conclusões e principais contribuições alcançadas no trabalho, juntamente com os indicativos de trabalhos futuros. Finalmente, as publicações realizadas durante o andamento dos trabalhos são relacionadas no Capítulo 6.

## 2 FUNDAMENTAÇÃO TEÓRICA

O grande volume de dados que trafegam nas redes de computadores tem causado um aumento da preocupação na detecção de conteúdos maliciosos, sendo o reconhecimento desses padrões um fator importante nas questões relacionadas à segurança em redes de computadores. Com o objetivo de tratar grandes volumes de dados que apresentam uma grande quantidade de amostras e características, a empregabilidade de redes neurais tem crescido nos últimos anos, apresentando resultados cada vez melhores em pesquisas relacionadas ao reconhecimento de padrões, principalmente quando o conjunto de amostras apresenta um grande número de características.

Neste capítulo, serão apresentadas as fundamentações teóricas importantes para o desenvolvimento deste trabalho. A Seção 2.1 aborda os conceitos sobre segurança em redes de computadores e suas preocupações no aprimoramento de técnicas existentes para detecções de anomalias focadas em *spam* em mensagens eletrônicas de e-mails. Já na Seção 2.2 é descrito o método de seleção de características empregado em uma das fases da condução dos experimentos para redução de dimensionalidade. Na Seção 2.3 têm-se os conceitos que cercam as RBMs e suas aplicações para este contexto. E por fim, na Seção 2.4, é apresentado o classificador OPF, o qual é utilizado na condução dos experimentos para treinamento e classificação, como uma forma de mensurar e comparar os resultados obtidos pelos métodos aplicados no aprendizado de características.

### 2.1 SEGURANÇA DE REDES DE COMPUTADORES

Segurança de redes de computadores refere-se a quaisquer atividades concebidas para proteger a usabilidade, confiabilidade, integridade e segurança de dados; tendo como alvo impedir que uma variedade de ameaças entrem e se espalhem pela rede. As ameaças são impostas através de intrusões a equipamentos conectados à rede, um dos objetivos mais comuns relacionados a estes ataques estão ligados principalmente à captura de informações de caráter sigiloso ou íntimo. A preocupação com segurança se deve ao crescimento das redes de computadores e ao número de pessoas interessadas em obter permissão a informações contidas em seu interior. Ferramentas e técnicas para interceptar tais invasões na rede estão sendo utilizadas em larga escala buscando aumentar a segurança dos sistemas de proteção, também conhecidos por Sis-

temas de Detecção de Intrusão, do inglês *Intrusion Detection Systems* (IDS) (ALLEN et al., 2000). Dessa forma, a área da segurança tem como objetivo preocupar-se com diversos métodos malignos de intrusões que possam ingressar na rede e provocar danos. Nas últimas décadas, houve um aumento na quantidade de diferentes invasões em redes de computadores, levando a uma preocupação com a identificação destes ataques e seu processo de detecção, sendo essa uma vertente do estudo proposto neste trabalho.

O conceito de detecção de anomalias pode ser aplicado em situações diversas, como a área de segurança em redes de computadores que vem sendo amplamente pesquisada e, como dito anteriormente, focando em técnicas para a detecção de tipos de anormalidades que possam colocar a segurança da rede em risco. Devido à diversificação dos ataques, bem como sua complexidade, os investimentos vêm sendo aumentados em estudos para a elaboração de sistemas de detecção de intrusões mais eficientes, os quais muitos deles são baseados em técnicas de inteligência artificial. Tais sistemas possuem como principal objetivo monitorar redes de computadores com o intuito de identificar e aprender novos tipos de ataques, automatizando, assim, o processo de monitoramento (WHITMAN; MATTORD, 2004).

A detecção de anomalias é amplamente utilizada para a detecção anormal ou incomum de padrões de dados. Para Das Kanishka Bhaduri (2011), dependendo da forma como as anomalias são definidas, diferentes algoritmos são desenvolvidos para encontrá-las a partir de um conjunto de dados, cada um com diferentes pressupostos e complexidades. Dentre as anomalias em redes de computadores, uma que tem grande destaque é o baseado em *spam*, que são foco dos experimentos conduzidos por este trabalho no aprendizado de características usando RBMs de forma não-supervisionada, voltado à detecção de *spams* em conteúdo de mensagens eletrônicas.

### 2.1.1 Detecção de Spam

O surgimento da internet trouxe amplos benefícios nas áreas de comunicação, entretenimento, compras, relações sociais, entre outras. Entretanto, várias ameaças começaram a surgir nesse cenário, levando pesquisadores a criarem ferramentas para lidar com esse foco. *Spam*, *malwares*, conteúdos maliciosos, *phishing*<sup>1</sup>, fraudes e falsos endereços de sites, do inglês *Uniform Resource Locator* (URL) são exemplos de ameaças. Em contrapartida, sistemas antivírus, *firewalls* e IDS são exemplos de ferramentas de combate às tais ameaças.

De maneira irrefutável, é realmente necessário o desenvolvimento de tais sistemas, visto que as ferramentas clássicas baseadas em assinatura são efetivas apenas contra 30-50% das

---

<sup>1</sup>*Phishing* é o ato criminoso de fisgar dados sensíveis de um usuário, como senhas, dados bancários, entre outros.

atuais ameaças de segurança e, além disso, é esperado que essa efetividade decresça com o tempo (IDC, 2012). Diante dessa situação, torna-se necessário buscar outros métodos que atinjam melhor desempenho contra as ameaças atuais e, nesse contexto, algoritmos de aprendizado de máquina podem ser utilizados. No atual cenário cibernético mundial, as principais ameaças vêm de *malwares* e *spam*, que também representam gastos desnecessários de tempo e dinheiro, além de ser um vetor para ataques de *phishing*.

Yang, Peng e Liu (2014) definem *spam* ou e-mail em massa não solicitado, ou ainda e-mail comercial não solicitado como a prática de enviar mensagens de e-mail não desejadas, geralmente com conteúdo comercial, em grande quantidade, para um número indiscriminado de pessoas. Neste contexto, vários problemas podem ser relacionados à esta prática, dentre os quais desperdício de largura de banda, espaço de armazenamento, sobrecarga do servidor, má fama associada ao e-mail como forma confiável de comunicação, *pishing* e conteúdo ofensivo são alguns exemplos (YANG; PENG; LIU, 2014; JATANA; SHARMA, 2014).

Ao longo dos anos, várias técnicas para filtragem de *spam* foram desenvolvidas. Kakade et al. (2014) apresentam um levantamento das principais delas, as quais foram divididas em quatro principais classes: técnicas baseadas em conteúdo, técnicas baseadas em listas, as híbridas ou outras técnicas de filtragem de *spam* e, por fim, técnicas baseadas em aprendizado de máquina.

As técnicas baseadas em conteúdo avaliam um e-mail através de suas palavras, frases ou anexos na intenção de determinar se o mesmo é *spam* ou legítimo (*ham*). Existem dois filtros que se valem dessa técnica: o baseados em palavras e o heurísticos:

- O filtro baseado em palavras bloqueia a mensagem baseado na “spamicidade” de algumas palavras, isto é, se um e-mail possui alguma palavra que frequentemente é utilizada em *spam*, ele é bloqueado. O problema dessa técnica é que se um filtro é configurado com palavras comuns, a taxa de falsos positivos aumenta consideravelmente;
- Por outro lado, os filtros heurísticos utilizam algum critério mais intuitivo do que simplesmente métricas técnicas. Pontuação (*score*) é, geralmente, um critério utilizado para classificar e-mail como *spam* ou *ham*. Para isso, mais pontos são dados a termos que mais frequentemente se encontram em *spams*, e termos geralmente usados em e-mails legítimos possuem baixa pontuação. Depois de avaliar as palavras presentes em um e-mail, uma pontuação é calculada e atribuída. Caso ultrapasse um limite preestabelecido, a mensagem é classificada como *spam*.

Técnicas baseadas em lista utilizam listas que mantêm os nomes de emissores legítimos e emissores de *spams*, conhecido como *spammers*. Uma mensagem é bloqueada, então, de



acordo com seu emissor. *Blacklist*, *Real Time BlackHole list*, *Whitelist* e *Greylist* são exemplos de técnicas baseadas em lista. Uma *Blacklist* contém uma lista de *spammers* e os *spams* são bloqueados através do nome de usuário ou do *Internet Protocol* (IP) presente nessa lista. A desvantagem é a taxa de falsos positivos no caso em que um *spammer* utiliza um IP ou usuário válido para mandar *spam*. *Real Time Blackhole list* é semelhante à citada acima, com a diferença que sua lista é mantida por terceiros, o que garante uma frequência maior de atualização. No entanto, uma organização pode ter menor controle sobre uma lista e casos de falsos positivos.

A classe de técnicas híbridas é caracterizada por utilizarem uma união ou mistura de técnicas, ou empregar algum outro meio para filtragem de *spam*. Um exemplo é o modelo *Bag-of-Words*, onde cada palavra no e-mail é listada em um documento e então associada com um índice. Assim, a frequência de ocorrência de cada palavra é utilizada como uma característica para treinar um classificador. As palavras mais comumente encontradas em *spam* vão para uma *bag*, que se comporta como uma classe, denominada *spam*. As outras palavras, ficam em uma *bag* de e-mails legítimos, assim sendo possível verificando em qual *bag* cada palavra desse e-mail se encontra, levando em consideração sua frequência.

Finalmente, as baseadas em aprendizado de máquina tendo como ponto forte sua base matemática e alta exatidão na classificação. Dentre elas, podem ser citadas, as técnicas baseadas em RNAs.

No contexto dos experimentos empregados neste trabalho foi utilizada técnica baseada no conteúdo, utilizando o modelo *Bag-of-Words* como representação, aplicando técnicas de RNAs baseadas em RBM para o aprendizado não-supervisionado de características visando a detecção de conteúdo malicioso, voltado à detecção de *spam* em mensagens eletrônicas.

## 2.2 REDUÇÃO DE DIMENSIONALIDADE

A aplicação do termo dimensionalidade está ligado diretamente a quantidade de características de uma representação de padrões, pode ser denominado também de dimensão do espaço de características. Quando se encontra um espaço de características que contenha as características mais relevantes a um determinado problema, isso pode impactar diretamente no processo de aprendizagem de um classificador, diminuindo o tempo deste processo e necessitando de um custo computacional menor, sendo estas as duas principais razões para que a dimensionalidade seja menor (JOACHIMS, 1998). Podem ser aplicados dois métodos de redução da dimensionalidade do espaço de características (HIRA; GILLIES, 2015), por:

- Extração: aplicação de uma transformação linear ou não linear sobre as variáveis originais

a fim de produzir um conjunto reduzido;

- Seleção: um subconjunto das variáveis originais é selecionado para o projeto do classificador.

A Análise de Componentes Principais, do inglês *Principal Component Analysis* (PCA) é um método clássico de projeção linear, baseado em extração. O método aplica uma transformação linear sobre um conjunto  $q$ -dimensional de dados de entrada e encontra um novo sistema de coordenadas de forma que a projeção de maior variância possível do conjunto de dados de entrada coincida com o primeiro eixo desse novo sistema, a de segunda maior variância, com o segundo eixo e assim sucessivamente, para os  $q$  novos eixos. Pode ser aplicada para obter uma redução de uma dimensão original  $q$  para uma dimensão reduzida  $u$  ( $u < q$ ) selecionando-se os  $u$  primeiros componentes principais de um determinado conjunto de dados e ignorando os demais (YU et al., 2009).

Nos problemas relacionados a detecção de *spam* os filtros baseados em seleção de características tem sido aplicados obtendo resultados interessantes na questão da diminuição da dimensionalidade, bem como no ato de encontrar características que melhor definem as classes em questão. No contexto deste trabalho, para a criação de um dicionário próprio para as bases de dados, foi aplicado um método baseado em seleção de características (MOHAMAD; SELAMAT, 2015).

### 2.2.1 Seleção de Características

Técnicas para seleção de características têm sido amplamente estudadas pela comunidade científica de reconhecimento de padrões e áreas afins, dado que o problema de encontrar o subconjunto das características que maximiza a taxa de acerto de uma técnica de classificação de padrões pode ser modelado como um problema de otimização (RODRIGUES, 2014). Tais técnicas são usadas para identificar quais termos são mais discriminantes para os algoritmos de classificação (JOACHIMS, 1998).

Na maioria das aplicações reais, as bases de dados possuem um grande número de características, muitas delas introduzidas para obter uma melhor representação do problema. Entretanto, grande parte destas características podem não ser relevantes, deste modo, tornando-se um problema comum para as aplicações reais. A seleção de características se refere a um processo no qual um espaço de dados é transformado em um espaço de características, de menor dimensão, mas que ainda retenha a maior parte da informação intrínseca dos dados; em outras palavras, o conjunto de dados sofre uma redução de dimensionalidade. Os métodos de seleção

de características tratam exatamente da escolha, dentre todos os atributos da base de dados, daqueles mais relevantes do ponto de vista da informação (DASH; LIU, 1997).

As questões relacionadas a dimensionalidade de características é agravado no caso da detecção de *spam* baseada em conteúdo, devido ao dicionário ser composto por palavras, onde cada uma representa uma característica. Dessa forma, uma base de dados pode ser formada por milhares de características, fazendo com que a aplicação de um método de seleção de característica seja importante para a montagem de um dicionário de palavras mais significativas.

O problema de seleção de características pode ser definido como o processo de encontrar um conjunto relevante de  $U$  características dentre as  $Q$  características originais, onde  $U \leq Q$ , para definir os dados (DASH; LIU, 1997). A seleção de características é um método utilizado para reduzir a dimensão do espaço de características, selecionando as palavras que são mais informativas para a classificação a ser realizada. É importante também que a seleção de características seja feita de forma automática. Podemos citar alguns algoritmos mais relevantes para este contexto, tais como, *Document Frequency Thresholding* (DF), *Mutual Information* (MI), *Information Gain* (IG), *Term Strength* (TS) e mais precisamente  $\chi^2$  Statistic ou Distribuição  $\chi^2$  (qui-quadrado). No que diz respeito a técnicas de seleção IG e  $\chi^2$ , tem sido mais eficaz na remoção de termos com índices menores de perda de precisão de classificação (ALMEIDA; YAMAKAMI; ALMEIDA, 2009).

### 2.2.1.1 Distribuição $\chi^2$

O processo de seleção representa a etapa que seleciona os termos mais representativos para as amostras, separando os termos mais relevantes em relação aos mais irrelevantes, sendo que os termos irrelevantes são eliminados do conjunto, obtendo uma redução de dimensionalidade. Existem duas maneiras de efetuar redução de dimensionalidade: Redução Local (realizada no conjunto de termos que ocorrem em cada categoria), ou Redução Global (realizada na base inteira). Neste trabalho, mais precisamente, é tratada a Redução Local com a técnica Distribuição  $\chi^2$  (qui-quadrado), do inglês  $\chi^2$  Statistic (SEBASTIANI, 2002), sendo caracterizada por ser uma técnica que mede o grau de independência entre um elemento em seu conjunto. Se  $y$  é uma característica e  $C$  um conjunto com duas classes, sendo elas neste caso, *spam* e *ham*, a distribuição  $\chi^2$  de uma característica  $y$  é dada pela Equação 1 (ASSIS, 2006):

$$\chi^2(y) = P(\textit{spam}) \cdot \chi^2(y, \textit{spam}) + P(\textit{ham}) \cdot \chi^2(y, \textit{ham}), \quad (1)$$

onde  $P(spam)$  e  $P(ham)$  são as probabilidades de ocorrência de e-mails *spam* e *ham*, respectivamente. A distribuição  $\chi^2$  para uma característica  $y$  em uma classe  $c$  é dada pela Equação 2:

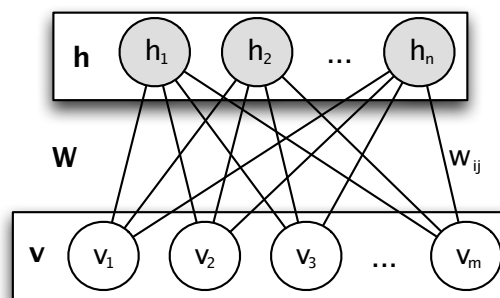
$$\chi^2(y, c) = \frac{Q \times (g \times q - u \times l)^2}{(g + u) \times (l + q) \times (g + l) \times (u + q)}, \quad (2)$$

onde  $g$  é o número de e-mails (dentro da classe  $c$ ), que contém a característica  $y$ ;  $l$  é o número de e-mails (dentro da classe  $\bar{c}$ ), que contém a característica  $y$ ;  $u$  é o número de e-mails (dentro da classe  $c$ ), que não contém a característica  $y$ ;  $q$  é o número de e-mails (dentro da classe  $\bar{c}$ ), que não contém a característica  $y$ ; e  $Q$  é o número total de e-mails na classe  $c$ . Após aplicado este processo, são escolhidas as características que apresentam valores mais altos, ou seja, que apresentam índices de significância maior para determinado problema (ASSIS, 2006).

### 2.3 MÁQUINAS DE BOLTZMANN RESTRITAS

Máquinas de Boltzmann Restritas são modelos de redes neurais estocásticas baseados em energia, sendo compostos por duas camadas de neurônios: uma camada de entrada (**visível**) e a outra **oculta**, onde o seu aprendizado é realizado de uma maneira **não-supervisionada**. Essencialmente, uma RBM é similar à clássica Máquina de Boltzmann (ACKLEY; HINTON; SEJNOWSKI, 1988), com a diferença que agora não são aceitas conexões entre neurônios de mesma camada<sup>2</sup>. A Figura 1 ilustra a arquitetura de uma RBM, a qual contém uma camada visível  $\mathbf{v}$  composta por  $m$  unidades, assim como uma camada oculta  $\mathbf{h}$  composta por  $n$  unidades. A matriz real-valorada  $\mathbf{W}_{m \times n}$  modela os pesos entre os neurônios das camada visível e oculta, onde  $w_{ij}$  corresponde ao peso entre a unidade visível  $v_i$  e a unidade oculta  $h_j$ .

Figura 1 - Arquitetura de uma RBM.



Fonte: Adaptado de Ackley, Hinton e Sejnowski (1988)

Primeiramente, RBMs foram projetadas utilizando somente unidades visíveis e ocultas binárias, as conhecidas Máquinas de Boltzmann Restritas de Bernoulli, do inglês *Bernoulli Res-*

<sup>2</sup>Na verdade, uma RBM pode ser vista como um grafo bipartido.

stricted Boltzmann Machines (BRBMs). Posteriormente, Welling et al. (WELLING; ROSENZVI; HINTON, 2005) evidenciaram outros tipos de unidades que podem ser utilizadas em uma RBM, tais como unidades Gaussianas e Binomiais<sup>3</sup>. Serão apresentados somente os conceitos básicos relacionados às BRBMs, os quais podem ser então generalizados para outros tipos de RBMs.

Sejam  $\mathbf{v}$  e  $\mathbf{h}$  as camadas visível e ocultas de uma RBM, respectivamente. Em outras palavras,  $\mathbf{v} \in \{0, 1\}^m$  e  $\mathbf{h} \in \{0, 1\}^n$ . A função de energia de uma RBM é dada pela Equação 3:

$$E(\mathbf{v}, \mathbf{h}) = - \sum_{i=1}^m a_i v_i - \sum_{j=1}^n b_j h_j - \sum_{i=1}^m \sum_{j=1}^n v_i h_j w_{ij}, \quad (3)$$

onde  $\mathbf{a}$  e  $\mathbf{b}$  correspondem às unidades de bias das camadas visível e oculta, respectivamente. A probabilidade de configuração  $(\mathbf{v}, \mathbf{h})$  é calculada de acordo com a Equação 4:

$$P(\mathbf{v}, \mathbf{h}) = \frac{e^{-E(\mathbf{v}, \mathbf{h})}}{\sum_{\mathbf{v}, \mathbf{h}} e^{-E(\mathbf{v}, \mathbf{h})}}, \quad (4)$$

onde o denominador da Equação 4 é um fator de normalização que representa todas as possíveis configurações envolvendo as camadas visível e oculta<sup>4</sup>. De uma maneira geral, o algoritmo de aprendizado de uma RBM objetiva estimar os parâmetros  $\mathbf{W}$ ,  $\mathbf{a}$  e  $\mathbf{b}$ , os quais podem ser otimizados por meio de um gradiente descendente estocástico no logaritmo da verossimilhança dos dados de treinamento. Assim, dada uma amostra de treinamento (unidade visível), a sua probabilidade é calculada sobre todas as possíveis unidades ocultas, conforme a Equação 5:

$$P(\mathbf{v}) = \frac{\sum_{\mathbf{h}} e^{-E(\mathbf{v}, \mathbf{h})}}{\sum_{\mathbf{v}, \mathbf{h}} e^{-E(\mathbf{v}, \mathbf{h})}}. \quad (5)$$

Com o intuito de atualizar os pesos e biases, é necessário calcular as seguintes derivativas, utilizando as Equações 6, 7 e 8:

$$\frac{\partial \log P(\mathbf{v})}{\partial w_{ij}} = E[h_j v_i]^{dado} - E[h_j v_i]^{modelo}, \quad (6)$$

$$\frac{\partial \log P(\mathbf{v})}{\partial a_i} = v_i - E[v_i]^{modelo}, \quad (7)$$

<sup>3</sup>Essas abordagens são também conhecidas como Máquinas de Boltzmann Restritas Gaussianas quando os neurônios da camada de entrada são modelados como sendo unidades Gaussianas.

<sup>4</sup>Note que essa normalização é extremamente custosa quando o número de unidades é muito grande.

$$\frac{\partial \log P(\mathbf{v})}{\partial b_j} = E[h_j]^{dado} - E[h_j]^{modelo}, \quad (8)$$

onde  $E[\cdot]$  corresponde à expectativa do argumento, e  $E[\cdot]^{dado}$  e  $E[\cdot]^{modelo}$  denotam as probabilidades dos dados e do modelo (dado reconstruído), respectivamente.

Em termos práticos,  $E[h_j v_i]^{dado}$  pode ser calculado considerando  $\mathbf{h}$  e  $\mathbf{v}$  como observado na Equação 9:

$$E[\mathbf{h}\mathbf{v}]^{dado} = P(\mathbf{h}|\mathbf{v})\mathbf{v}^T, \quad (9)$$

onde  $P(\mathbf{h}|\mathbf{v})$  corresponde à probabilidade de obter  $\mathbf{h}$  considerando um dado de treinamento  $\mathbf{v}$ , de acordo com a Equação 10:

$$P(h_j = 1|\mathbf{v}) = \sigma \left( \sum_{i=1}^m w_{ij} v_i + b_j \right), \quad (10)$$

onde  $\sigma(\cdot)$  denota a função sigmóide-logística<sup>5</sup>.

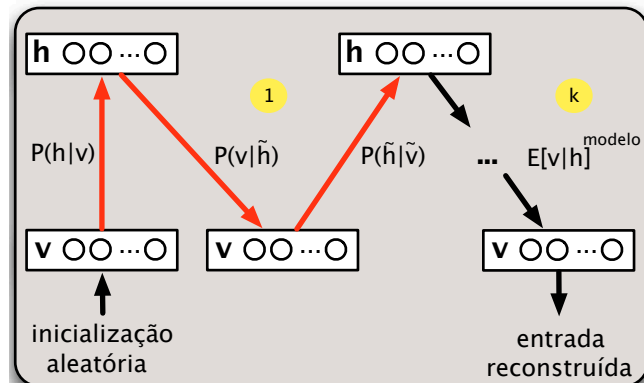
A grande questão agora consiste em obter  $E[h_j]^{modelo}$ , o qual é o modelo aprendido pelo sistema. Uma estratégia possível é utilizar amostragens de Gibbs sucessivas por meio de uma inicialização aleatória do estado inicial (unidade visível) até algum critério de convergência ser atingido. A amostragem de Gibbs consiste, basicamente, em atualizar as unidades ocultas utilizando a Equação 10, e depois atualizar as unidades visíveis utilizando  $P(\mathbf{v}|\mathbf{h})$ , dada pela Equação 11:

$$P(v_i = 1|\mathbf{h}) = \sigma \left( \sum_{j=1}^n w_{ij} h_j + a_i \right), \quad (11)$$

e atualizando novamente as unidades ocultas por meio da Equação 10. Assim, é possível obter uma estimativa do modelo  $E[\mathbf{h}\mathbf{v}]^{modelo}$  inicializando as unidades visíveis com valores aleatórios e depois executar uma amostragem de Gibbs utilizando  $k$  iterações, conforme ilustra a Figura 2. Com o intuito de facilitar a explicação, o exemplo emprega  $P(\mathbf{v}|\tilde{\mathbf{h}})$  ao invés de  $P(\mathbf{v}|\mathbf{h})$ , e  $P(\tilde{\mathbf{h}}|\tilde{\mathbf{v}})$  ao invés de  $P(\mathbf{h}|\mathbf{v})$ . Essencialmente, essas terminologias possuem mesmo significado, mas  $P(\mathbf{v}|\tilde{\mathbf{h}})$  é utilizado aqui para denotar que a unidade visível  $\mathbf{v}$  será reconstruída usando  $\tilde{\mathbf{h}}$ , a qual foi obtida por meio de  $P(\mathbf{h}|\mathbf{v})$ . O mesmo ocorre com  $P(\tilde{\mathbf{h}}|\tilde{\mathbf{v}})$ , que reconstrói  $\tilde{\mathbf{h}}$  utilizando  $\tilde{\mathbf{v}}$ , que foi obtida por meio de  $P(\mathbf{v}|\tilde{\mathbf{h}})$ .

<sup>5</sup>A função sigmóide-logística pode ser calculada pela seguinte Equação:  $\sigma(g) = 1/(1 + \exp(-g))$ .

Figura 2 - Amostragem de Gibbs para estimar o modelo de reconstrução dos dados.



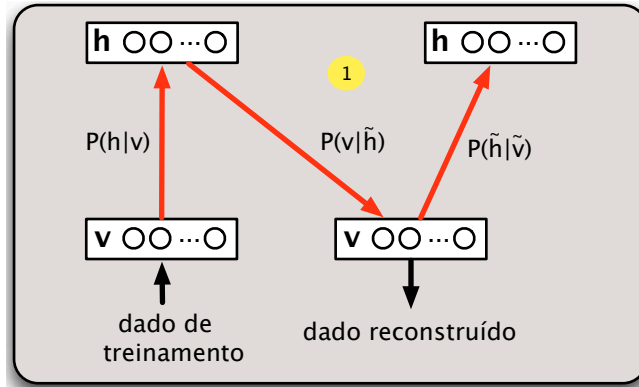
Fonte: Adaptado de Ackley, Hinton e Sejnowski (1988)

Entretanto, o procedimento apresentado na Figura 2 é caro computacionalmente, dado que é esperada uma boa reconstrução do modelo quando  $k \rightarrow +\infty$ . Dessa forma, alguns trabalhos apresentaram alternativas à amostragem de Gibbs, sendo alguns deles o método da **Divergência Contrastiva**, do inglês *Contrastive Divergence* (CD) (HINTON, 2002), **Divergência Contrastiva Persistente**, do inglês *Persistent Contrastive Divergence* (PCD) (TIELEMAN, 2008) e **Divergência Contrastiva Persistente Rápida**, do inglês *Fast Persistent Contrastive Divergence* (FPCD) (TIELEMAN; HINTON, 2009). Todas essas abordagens são baseadas em cadeias de Markov, sendo basicamente utilizadas para amostrar dados dessas cadeias, ou seja, realizar o processo de treinamento de uma RBM. Este processo pode ser conduzido por três técnicas de aprendizado, baseadas em: Divergência Contrastiva, Divergência Contrastiva Persistente e Divergência Contrastiva Persistente Rápida.

### 2.3.1 Divergência Contrastiva

Hinton (2002) introduziu uma metodologia mais rápida para o cálculo de  $E[\mathbf{h}\mathbf{v}]^{modelo}$  baseado na ideia de divergência contrastiva. Basicamente, a ideia é inicializar as unidades visíveis com amostras de treinamento e estimar os estados das unidades ocultas utilizando a Equação 10, e então calcular os estados da unidade visível (passo de reconstrução) por meio da Equação 11. Em suma, isso seria equivalente a executar a amostragem de Gibbs usando  $k = 1$  com a cadeia inicializada com uma amostra de treinamento. A Figura 3 ilustra essa abordagem.

Figura 3 - Ilustração do método baseado em divergência contrastiva.



Fonte: Adaptado de Ackley, Hinton e Sejnowski (1988)

A noção de “cadeia de Markov” em uma RBM, basicamente, diz respeito ao conjunto de dados reconstruídos durante o processo de amostragem, os quais são calculados pela Equação 11. Seja  $P^k(\mathbf{v}|\mathbf{h})$  a estimativa do dado de entrada  $\mathbf{v}$  na  $k$ -ésima iteração de algum processo de amostragem, para um processo com  $k = 3$  iterações, neste caso, tem-se a seguinte cadeia:  $P^1(\mathbf{v}|\mathbf{h}) \rightarrow P^2(\mathbf{v}|\mathbf{h}) \rightarrow P^3(\mathbf{v}|\mathbf{h})$ , onde  $P^1(\mathbf{v}|\mathbf{h})$  denota a primeira etapa de reconstrução (seria o modelo estimado pelo método CD), e  $P^3(\mathbf{v}|\mathbf{h})$  denota o último elemento dessa cadeia.

Com base na ideia da convergência contrastiva,  $E[\mathbf{h}\mathbf{v}]^{modelo}$  pode ser calculado em conformidade com a equação 12:

$$E[\mathbf{h}\mathbf{v}]^{modelo} = P(\tilde{\mathbf{h}}|\tilde{\mathbf{v}})\tilde{\mathbf{v}}^T. \quad (12)$$

Portanto, a Equação 12 leva a uma regra de aprendizado para a atualização da matriz de pesos  $\mathbf{W}$ , como observado na Equação 13:

$$\begin{aligned} \mathbf{W}^{t+1} &= \mathbf{W}^t + \eta(E[\mathbf{h}\mathbf{v}]^{dado} - E[\mathbf{h}\mathbf{v}]^{modelo}) \\ &= \mathbf{W}^t + \eta(P(\mathbf{h}|\mathbf{v})\mathbf{v}^T - P(\tilde{\mathbf{h}}|\tilde{\mathbf{v}})\tilde{\mathbf{v}}^T), \end{aligned} \quad (13)$$

onde  $\mathbf{W}^t$  corresponde à matriz de pesos na iteração  $t$ , e  $\eta$  denota a taxa de aprendizado. Adicionalmente, empregam-se as Equações 14 e 15 para atualizar o bias das unidades visível e ocultas:

$$\mathbf{a}^{t+1} = \mathbf{a}^t + \eta(\mathbf{v} - E[\mathbf{v}]^{modelo})$$



$$= \mathbf{a}^t + \eta(\mathbf{v} - \tilde{\mathbf{v}}), \quad (14)$$

e

$$\begin{aligned} \mathbf{b}^{t+1} &= \mathbf{b}^t + \eta(E[\mathbf{h}]^{dado} - E[\mathbf{h}]^{modelo}) \\ &= \mathbf{b}^t + \eta(P(\mathbf{h}|\mathbf{v}) - P(\tilde{\mathbf{h}}|\tilde{\mathbf{v}})), \end{aligned} \quad (15)$$

onde  $\mathbf{a}^t$  e  $\mathbf{b}^t$  denotam os vetores de bias das unidades visíveis e ocultas, respectivamente. Assim, as Equações 13, 14 e 15 correspondem à formulação básica para atualização dos pesos de uma RBM.

Posteriormente, Hinton (2012) introduziu um parâmetro de decaimento de peso  $\lambda$ , o qual penaliza os pesos com grande magnitude<sup>6</sup>, bem como o parâmetro de momento  $\alpha$  para controlar possíveis alterações durante o processo de aprendizado. Portanto, as Equações 13, 14 e 15 podem ser re-escritas como segue<sup>7</sup>:

$$\mathbf{W}^{t+1} = \mathbf{W}^t + \underbrace{\eta(P(\mathbf{h}|\mathbf{v})\mathbf{v}^T - P(\tilde{\mathbf{h}}|\tilde{\mathbf{v}})\tilde{\mathbf{v}}^T)}_{=\Delta\mathbf{W}^t} - \lambda\mathbf{W}^t + \alpha\Delta\mathbf{W}^{t-1}, \quad (16)$$

$$\mathbf{a}^{t+1} = \mathbf{a}^t + \underbrace{\eta(\mathbf{v} - \tilde{\mathbf{v}})}_{=\Delta\mathbf{a}^t} + \alpha\Delta\mathbf{a}^{t-1} \quad (17)$$

e

$$\mathbf{b}^{t+1} = \mathbf{b}^t + \underbrace{\eta(P(\mathbf{h}|\mathbf{v}) - P(\tilde{\mathbf{h}}|\tilde{\mathbf{v}}))}_{=\Delta\mathbf{b}^t} + \alpha\Delta\mathbf{b}^{t-1}. \quad (18)$$

### 2.3.2 Divergência Contrastiva Persistente

Um dos principais problemas relacionados à abordagem de amostragem por divergência contrastiva diz respeito à quantidade de iterações utilizadas para aproximar a estimativa do modelo de seu valor real (valor de  $k$  na Figura 2). Muito embora a abordagem proposta por Hinton (2002) utilize  $k = 1$  e, na prática, pode funcionar muito bem, outros valores para essa variável  $k$  também podem ser estabelecidos<sup>8</sup> (CARREIRA-PERPINAN; HINTON, 2005).

Não obstante a divergência contrastiva seja uma boa aproximação da verossimilhança do

<sup>6</sup>Os pesos podem aumentar consideravelmente durante o processo de convergência do algoritmo.

<sup>7</sup>Note que quando  $\lambda = 0$  e  $\alpha = 0$ , tem-se a formulação básica mencionada anteriormente.

<sup>8</sup>Frequentemente, o método da divergência contrastiva com apenas uma iteração é denominado de CD-1.

gradiente, isto é, seja uma boa aproximação do modelo de dados de entrada quando  $k \rightarrow \infty$ , a sua convergência pode não ser muito boa quando a cadeia de Markov possui uma “mistura baixa”<sup>9</sup>. Além disso, o método da divergência contrastiva possui uma taxa de convergência muito boa apenas nas iterações iniciais, ficando mais lenta à medida que as iterações vão aumentando, obrigando o uso de técnicas de decaimento nos parâmetros da abordagem a ser empregada (taxa de aprendizagem de uma RBM, conforme apresentado na Equações 16, 17 e 18, por exemplo).

Desta forma, uma alternativa interessante seria a utilização de um valor maior para  $k$  no método da divergência contrastiva, usualmente denominado CD- $k$ . Entretanto, o maior problema dessa abordagem diz respeito ao custo computacional, dado que mais iterações serão agora utilizadas para aproximar o modelo dos dados de entrada. Baseado nessas premissas, Tieleman (2008) propôs o método da Divergência Contrastiva Persistente, o qual objetiva uma aproximação do modelo muito semelhante ao CD- $k$ , mas com um custo computacional inferior. No método CD-1, cada amostra de treinamento é então utilizada para inicializar a RBM e, após uma iteração do processo de amostragem de Gibbs, obtém-se então o modelo reconstruído, conforme apresentado na Figura 3. Após todas as amostras do conjunto de treinamento serem apresentadas à RBM, tem-se então uma iteração (época) do processo total de aprendizado dos pesos. Para a segunda época, todo esse processo é repetido, cada amostra de treinamento é utilizada para inicializar a RBM, ou seja, a cadeia de Markov de cada amostra é reinicializada. Já a técnica PCD proposta por Tieleman visa uma aproximação do modelo “ideal” dos dados obtido com CD- $k$  (com  $k \rightarrow \infty$ ) e, para obter essa estimativa, essa técnica propõe não reinicializar a cadeia de cada amostra a cada nova época, e sim utilizar o último elemento dessa cadeia na iteração anterior como sendo o novo valor para essa cadeia de Markov. Assim sendo, à medida que as épocas vão avançando, o método vai se aproximando de um modelo similar ao obtido quando empregamos CD- $k$ .

### 2.3.3 Divergência Contrastiva Persistente Rápida

Muito embora a técnica da Divergência Contrastiva Persistente apresentada na seção anterior possa ter um desempenho melhor do que a Divergência Contrastiva tradicional em várias situações, alguns problemas ainda podem ocorrer. De acordo com Tieleman e Hinton (2009), se a cadeia persistente for amostrada a partir da distribuição atual do modelo  $\Theta$ , onde  $\Theta = \{\mathbf{W}, \mathbf{a}, \mathbf{b}, \eta, \lambda, \alpha\}$  corresponde ao conjunto de parâmetros da RBM, as atualizações dos pesos (i.e.,  $\mathbf{W}$ ,  $\mathbf{a}$  e  $\mathbf{b}$ ) seguem a direção do gradiente do logaritmo da verossimilhança, o que é muito bom, pois aproxima-se do “resultado ideal”. Entretanto, se a cadeia for amostrada a

---

<sup>9</sup>O termo mistura é frequentemente utilizado por artigos que tratam de métodos de amostragem em cadeias de Markov como *mixing*, denotando a qualidade do processo de convergência dessa cadeia.

partir de uma distribuição diferente, diga-se  $\bar{\Theta}$ , existe um “termo adicional” no momento de atualização dos pesos que move a distribuição do modelo para uma outra diferente de  $\bar{\Theta}$ , o que não é bom para a RBM<sup>10</sup>.

Assim sendo, Tieleman e Hinton (2009) propuseram o método da Divergência Contrastiva Persistente Rápida, onde um conjunto de “pesos rápidos” é adicionado no intuito de forçar uma RBM a “desaprender” quando a mesma está sendo atraída para ótimos locais na função de energia. Basicamente, os pesos tradicionais, os quais são usados normalmente nas técnicas CD e PCD, são utilizados para o aprendizado da RBM, enquanto que o novo conjunto de pesos introduzido é então utilizado para fazer com que a RBM “desaprenda” em caso de uma convergência fraca do método. Dessa forma, enfatizam a grande importância da sinergia entre aprendizado e convergência de uma RBM.

As características geradas pelo processo de aprendizado não-supervisionado das RBMs são passadas para o classificador OPF.

## 2.4 FLORESTA DE CAMINHOS ÓTIMOS

O classificador OPF faz uso da Transformada Imagem Floresta, do inglês *Image Foresting Transform* (IFT) (FALCÃO; STOLFI; LOTUFO, 2004), que deriva uma imagem em um grafo e torna possível o cálculo de uma floresta de caminhos de custo mínimo. Para isso, considere a abstração de que uma imagem seja um grafo, onde os nós são pixels e as arestas são definidas por uma relação de adjacência  $A$  entre os nós. Um caminho nesse grafo é uma sequência de amostras  $\pi_{s_x} = \langle s_1, s_2, \dots, s_x \rangle$ , onde  $(s_o, s_{o+1}) \in A$  para  $1 \leq o \leq x - 1$ . A cada caminho  $\pi_o$  é associado uma função de valor de caminho  $f$ , denotada  $f(\pi_s)$ . Como exemplo de uma função de valor de caminho, pode-se citar  $f_{max}$ , definida da seguinte maneira:

$$\begin{aligned} f_{max}(\langle s \rangle) &= \begin{cases} 0 & \text{se } s \in S, \\ +\infty & \text{caso contrário} \end{cases} \\ f_{max}(\pi \cdot \langle s, t \rangle) &= \max\{f_{max}(\pi), d(s, t)\}, \end{aligned} \quad (19)$$

onde  $\pi \cdot \langle s, t \rangle$  é a concatenação do caminho  $\pi_s$  com término em  $s$  e o arco  $(s, t)$ ,  $d(s, t)$  mede a dissimilaridade entre nós adjacentes e  $f_{max}(\pi_s)$  computa a distância máxima entre amostras adjacentes em  $\pi_s$ .

A abordagem supervisionada para classificação de padrões trata as amostras como nós de

---

<sup>10</sup>Esse termo adicional diz respeito ao gradiente de  $KL(\bar{\Theta}||\Theta)$ , onde  $KL$  corresponde à Divergência de Kullback-Leibler, que nada mais é do que a diferença entre as distribuições de probabilidades  $\bar{\Theta}$  e  $\Theta$ .

um grafo e os arcos são definidos por uma relação de adjacência e ponderados por alguma métrica de distância aplicada a seus vetores de atributos. Para tratar dessa abordagem, dois métodos são utilizados, os quais se diferem tanto na relação de adjacência e função de valor de caminho utilizadas quanto na maneira de encontrar os protótipos. A primeira usa como relação de adjacência o grafo completo e busca como protótipos amostras que pertencem à interseção entre as classes no conjunto de treinamento (PAPA; FALCÃO; SUZUKI, 2009; PAPA et al., 2012). A outra utiliza um grafo  $k$ -nn e encontra os protótipos como sendo os máximos regionais ou amostras de cada classe na junção entre as densidades de probabilidade (ROCHA; CAPPABIANCO; FALCÃO, 2009). O presente trabalho contempla a abordagem com grafo completo por ser mais rápido, dado que este não possui parâmetros.

### 2.4.1 Grafo Completo

Neste método, as amostras são representadas por nós de um grafo completo, isto é, todas as amostras são conectadas entre si. Os elementos mais representativos de cada classe do conjunto de treinamento, isto é, os protótipos, são escolhidos como sendo os elementos pertencentes às regiões de fronteira entre as classes, os quais participam de um processo de competição disputando outras amostras oferecendo-lhes caminhos de menor custo e seus respectivos rótulos. Ao final desse processo, obtemos um conjunto de treinamento particionado em árvores de caminhos ótimos, sendo que a união das mesmas nos remete a uma floresta de caminhos ótimos.

Para compreender o funcionamento deste método, considere  $Z$  uma base de dados  $\lambda$ -rotulada,  $Z_1$  e  $Z_2$  os conjuntos de treinamento e teste, respectivamente, tal que  $Z = Z_1 \cup Z_2$  e  $\lambda(s)$  uma função que associa o rótulo correto  $o, o = 1, 2, \dots, cl$  da classe  $o$  a qualquer amostra  $s \in Z_1 \cup Z_2$ .

Seja  $S \subset Z_1$  um conjunto de protótipos de todas as classes,  $\vec{A}$  um algoritmo que extrai  $x$  atributos de qualquer amostra  $s \in Z_1 \cup Z_2$  e retorna um vetor de atributos  $\vec{A}(s) \in \mathfrak{R}^x$ . O algoritmo baseado em OPF associa um caminho ótimo  $p^*(s)$  de um elemento do conjunto de protótipos  $S$  a toda amostra  $s \in Z_1$ , formando uma floresta de caminhos ótimos  $p$ , ou uma marca *nil* quando  $s \in S$ .

Seja  $R(s) \in S$  a raiz de  $p^*(s)$  a qual pode ser alcançada por  $p(s)$ , sendo este o predecessor de  $s$ . O algoritmo computa, para cada  $s \in Z_1$ , o custo  $H(s)$  de  $p^*(s)$ , o rótulo  $L(s) = \lambda(R(s))$  e seu predecessor  $p(s)$  da forma apresentada pelo Algoritmo da Tabela 1.

As linhas 1 e 2 inicializam, para todas as amostras  $s \in Z_1$  e protótipos  $s \in S$ , os mapas de valores de custo de caminho e de rótulos. Os predecessores são inicializados como *nil*. A cada iteração, um caminho de custo ótimo  $H(s)$  é obtido em  $p$ , sendo este uma floresta de

caminhos ótimos. A política FIFO (*first in first out*) evita empates, pois quando dois caminhos atingem uma determinada amostra  $s$  com o mesmo custo mínimo, essa amostra  $S$  é associada ao primeiro caminho que o atingiu. Das linhas 5-10, o algoritmo verifica se o caminho que atinge uma amostra adjacente  $t$  através de  $s$  possui menor custo que o caminho que termina em  $t$ . Se for, e caso  $H(t) \neq +\infty$ ,  $s$  conquista  $t$ . Assim,  $s$  se torna predecessor de  $t$ ,  $t$  recebe o rótulo de  $s$  e o valor do custo é atualizado.

Tabela 1 - Algoritmo Classificador Supervisionado baseado em Floresta de Caminhos Ótimos usando grafo completo.

---

**Algoritmo 1:** CLASSIFICADOR SUPERVISIONADO BASEADO EM FLORESTA DE CAMINHOS ÓTIMOS USANDO GRAFO COMPLETO

---

**Entrada:** Um conjunto de treinamento  $Z_1$   $\lambda$ -rotulado, protótipos  $S \subset Z_1$  e o par  $(v, d)$  para vetor de características e cálculo das distâncias.

**Saída:** Floresta de caminhos ótimos  $P$ , mapa de valores de custo de caminhos  $H$  e mapa de rótulos  $L$ .

```

1 início
2   para todo  $s \in Z_1$  faça
3     |  $p(s) = nil$  e  $H(s) = +\infty$ .
4   fim
5   para todo  $s \in S$  faça
6     |  $H(s) = 0$ ,  $p(s) = nil$ ,  $L(s) = \lambda(s)$  e insira  $s$  em  $Qx$ .
7   fim
8   repita
9     | Remova de  $Qx$  uma instância  $s$  tal que  $H(s)$  é mínimo.
10    repita
11      |  $cst = \max\{V(s), d(s, t)\}$ 
12      se  $cst < V(t)$  então
13        | se  $H(t) \neq +\infty$  então
14          | remova  $t$  de  $Qx$ .
15        fim
16        |  $P(t) = s$ ,  $L(t) = L(s)$ ,  $H(t) = cst$  e Insira  $t$  em  $Qx$ .
17      fim
18    até  $t \in Z_1$  tal que  $s \neq t$  e  $H(t) > V(s)$ ;
19  até  $Qx$  é não vazia;
20 fim

```

---

Fonte: Papa, Falcão e Suzuki (2009).

## 2.4.2 Algoritmo de Treinamento

A fase de treinamento consiste em encontrar o conjunto  $S$  de protótipos que serão estimados nas regiões de sobreposição de amostras e nas fronteiras entre as classes, visto que são regiões muito susceptíveis a erros de classificação.

Computando uma Árvore de Espalhamento Mínimo, do inglês *Minimum Spanning Tree* (MST), no grafo completo  $(Z_1, A)$ , obtemos um grafo conexo acíclico cujos nós são todas as amostras em  $Z_1$  e os arcos são não direcionados e ponderados. Seus pesos são dados pela distância  $d$  entre os vetores de atributos de amostras adjacentes. Os protótipos a serem escolhidos são os elementos conectados na MST com diferentes rótulos em  $Z_1$ , isto é, elementos mais próximos de classes diferentes. Removendo-se os arcos entre classes diferentes, tais amostras adjacentes tornam-se protótipos em  $S$  e o algoritmo representado na Tabela 1 pode computar uma floresta de caminhos ótimos em  $Z_1$ .

### 2.4.3 Método de Classificação

Para qualquer amostra  $t \in Z_2$ , consideram-se todos os arcos conectando  $t$  com amostras  $s \in Z_1$ , tornando  $t$  como se fosse parte do grafo original. Considerando todos os possíveis caminhos entre  $S$  e  $t$ , deseja-se encontrar o caminho ótimo  $P^*(t)$  de  $S$  até  $t$  com a classe  $\lambda(R(t))$  de seu protótipo  $R(t) \in S$  mais fortemente conexo. Este caminho pode ser identificado incrementalmente, avaliando o valor do custo ótimo  $H(t)$  conforme a Equação 20:

$$H(t) = \min\{\max\{H(s), d(s, t)\}\}, \forall s \in Z_1. \quad (20)$$

Nota-se que os classificadores baseados em OPF utilizam a força de conectividade entre as amostras para a classificação dos dados, ou seja, não são algoritmos baseados em distância apenas.

### 3 MATERIAIS E MÉTODOS

As RBMs têm sido amplamente utilizadas para o aprendizado não-supervisionado de características nos últimos anos (HINTON, 2002). São redes neurais estocásticas que visam prever algumas variáveis latentes, também conhecidas como neurônios ocultos. Portanto, com alguns dados rotulados, a ideia é reconstruí-los com base em previsões sobre a camada oculta. O processo de reconstrução é normalmente realizado por várias amostragens em uma cadeia de Markov (*Markov Chain*), sendo a última amostragem nessa cadeia utilizada como a reconstrução final do dado de entrada. Este processo também acaba por modificar os pesos de ligação entre a entrada e os neurônios ocultos, e após o processo de aprendizagem, pode ser utilizado um conjunto de teste para prever as variáveis latentes. No entanto, um dos principais problemas relacionados às RBMs são as preocupações com o ajuste fino de seus parâmetros, uma vez que tais técnicas são muito sensíveis à sua seleção de parâmetros. Papa et al. (PAPA et al., 2015a, 2015b; PAPA; SCHEIRER; COX, 2015) foram um dos primeiros a empregar técnicas de meta-heurísticas nesse contexto, alcançando resultados promissores. Basicamente, o processo é modelar o problema de ajuste fino de parâmetros como uma tarefa de otimização. Dessa forma, com o objetivo de melhorar a detecção de conteúdo malicioso pelo método do aprendizado de características através da RBM, na etapa de aprendizado é realizada uma otimização de quatro parâmetros: a taxa de aprendizado (*learning rate*)  $\eta$ , o decaimento dos pesos (*weight decay*)  $\lambda$ , o parâmetro de penalidade (*penalty parameter*)  $\alpha$  e o número de unidades ocultas (*hidden units*)  $n$ ; através da aplicação de algoritmos meta-heurísticos de otimização baseados em técnicas de Busca Harmônica. As técnicas aplicadas são:

- Busca Harmônica Discreta (GEEM; KIM; LOGANATHAN, 2001): visa modelar o problema da função de minimização com base na maneira como músicos criam suas canções com ótimas harmonias. Essa técnica emprega dois parâmetros para resolver este problema, sendo a Taxa de Consideração de Memória Harmônica, do inglês *Harmony Memory Considering Rate* (HMRC), responsável pela criação de novas soluções com base em experiência anterior do reprodutor de música, e a Taxa de Ajuste de Afinação, do inglês *Pitch Adjusting Rate* (PAR)<sup>1</sup>;
- Busca Harmônica Aprimorada, do inglês *Improved Harmony Search* (IHS) (MAHDAVI;

---

<sup>1</sup>Responsável pela aplicação de alguma pequena perturbação na solução criada com HMRC, a fim de evitar as armadilhas de ótimos locais.

FESANGHARY; DAMANGIR, 2007): essa abordagem utiliza valores dinâmicos para ambas variáveis HMCR e PAR, que são atualizados a cada iteração com os novos valores que estão dentro do intervalo  $[HMCR_{min}, HMCR_{max}]$  e  $[PAR_{min}, PAR_{max}]$ , respectivamente. Para o cálculo de PAR é utilizada a variável largura de banda (*bandwidth*)  $\rho$ , e seus valores devem estar entre  $[\rho_{min}, \rho_{max}]$ ;

- Busca Harmônica Livre de Parametrização, do inglês *Parameter Setting-Free Harmony Search* (PSF-HS) (GEEM; SIM, 2010): proposta a fim de evitar a etapa de parâmetros de ajuste fino com relação a HMCR, PAR e  $\rho$ . A ideia é obter novos valores HMCR e PAR com base em cálculos anteriores de tais variáveis, sempre que uma nova solução é criada.

Na condução dos experimentos na configuração das RBMs foram empregados 10 agentes sobre 50 interações para a convergência considerando todas as técnicas. A Tabela 2 apresenta a configuração dos parâmetros de cada técnica de otimização baseada em HS<sup>2</sup>. Também foi aplicado  $T = 100$  como o número de épocas do procedimento de aprendizado dos pesos.

Tabela 2 - Configuração dos parâmetros meta-heurísticos de otimização baseados em técnicas de Busca Harmônica.

Técnica	Parâmetros
HS	$HMCR = 0.7, PAR = 0.7, \rho = 0.1$
IHS	$HMCR = 0.7, PAR_{MIN} = 0.1$ $PAR_{MAX} = 1.0, \rho_{MIN} = 0.1$ $\rho_{MAX} = 0.5$

Fonte: Dados da pesquisa do autor.

Baseados nas técnicas apresentadas e dentro da linha de detecção de anomalias, foram conduzidos experimentos usando o aprendizado não-supervisionado de características através do emprego das RBMs, tendo seus parâmetros refinados via a aplicação de algoritmos meta-heurísticos de otimização baseados em técnicas de HS e explorando a questão de detecção de *spam* em mensagens eletrônicas por seu conteúdo (dicionário de palavras). Parte dos experimentos deste trabalho foram publicados no artigo apresentado pelo autor, Silva et al. (SILVA et al., 2016).

A seguir, são descritas as bases de dados utilizadas nos experimentos e os protocolos aplicados em dois conjuntos de experimentos distintos, denominados de EXP1 e EXP2, respectivamente.

<sup>2</sup>Note que esses valores foram definidos empiricamente.



### 3.1 BASES DE DADOS

Para condução dos experimentos, foram utilizadas bases de dados públicas bem conhecidas na área de detecção de *spam* em e-mails:

- SPAMBASE<sup>3</sup>: esta base de dados está disponível na *UCI Machine Learning Repository* (LICHMAN, 2013), sendo constituída de 4.601 instâncias com 39,4% das amostras classificadas como mensagens de *spam*. Nessa base, cada instância é representada por 57 características, sendo 48 compostas por palavras referenciadas por frequência, que representam a frequência da ocorrência de cada palavra baseado no dicionário preestabelecido por seus criadores;
- LINGSPAM<sup>4</sup>: esta base de dados contém mensagens de e-mails de *spam* e *ham* coletadas via mensagens legítimas enviadas para uma lista de linguística, e consiste de 2.893 mensagens de e-mail, das quais 2.412 são rotuladas como *ham* e 481 como *spam* (ANDROUSOPOULOS et al., 2000);
- CSDMC<sup>5</sup>: esta base de dados é composta de uma seleção de mensagens, dividida em duas partes, uma de treinamento e outra de teste. O conjunto de treinamento contém 4.327 mensagens, sendo 2.949 classificadas como *ham* e 1.378 classificadas como *spam*. A taxa de *spam* nesta base de dados é de 32%. Essa base é relativamente nova e ainda é pouco explorada (GROUP, 2010).

Não houve a preocupação da utilização de bases mais atuais em virtude de estar analisando a aplicação das técnicas para o aprendizado não-supervisionado de características propostas neste trabalho, processo o qual não é impactado pela época que os dados foram coletados.

Para todas as bases apresentadas anteriormente, o conjunto de dados total que as compõem foram estratificados utilizando o método *holdout*, criando assim a divisão entre dados para treinamento e dados para teste<sup>6</sup>. A fim de estabelecer a acurácia média, foram gerados 10 conjuntos de arquivos de treinamento e teste. Como o intuito é validar e mensurar a aplicação das técnicas propostas neste trabalho para o aprendizado de características mais relevantes, utilizou-se, portanto, um único método de medição para acurácia obtida. No caso, foi aplicado o classificador OPF, que apresenta sua própria métrica de acurácia (PAPA; FALCÃO; SUZUKI, 2009).

<sup>3</sup><https://archive.ics.uci.edu/ml/datasets/Spambase>

<sup>4</sup><http://csmining.org/index.php/ling-spam-datasets.html>

<sup>5</sup><http://csmining.org/index.php/spam-email-datasets-.html>

<sup>6</sup>Este método consiste em dividir o conjunto total de dados em dois subconjuntos mutuamente exclusivos, um para treinamento e outro para teste.

As amostras das bases de dados já encontram-se previamente rotuladas entre *ham* ou *spam*, rótulos utilizados apenas na aplicação do classificador para medir e comparar a acurácia entre as amostras com as características originais e com as características aprendidas frente a um *pipeline* tradicional de treinamento e classificação, no caso aplicando o classificador OPF. É importante destacar que os atributos das bases de dados foram binarizados (0 e 1) de acordo com a falta ou presença de frequência da palavra (característica) na mensagem de e-mail, processo necessário uma vez que está sendo usada RBM tradicional.

Para fins de comparação da aplicação do método proposto, foram conduzidos dois experimentos, denominados de EXP1 e EXP2, cada um com duas variações de parâmetros denominadas CONF1 e CONF2 com o objetivo de avaliar diferentes intervalos de valores a respeito dos parâmetros da RBM e seus impactos nos resultados. Posteriormente ao processo de aprendizado de características empregado pelas RBMs, foi aplicado o classificador OPF para avaliação da eficiência do processo de aprendizado realizado frente aos dados originais.

### 3.2 PROTOCOLO DOS EXPERIMENTOS: EXP1

Nos experimentos denominados de EXP1, o dicionário de palavras predefinido na base de dados SPAMBASE foi pré-processado nas demais bases de dados: LINGSPAM e CSDMC com o objetivo de avaliar o impacto de um dicionário de palavras pré-estabelecido por uma base em outros conjuntos de dados. No caso, 48 palavras compõem o dicionário, resultando assim compondo 48 características. O processo *holdout* aplicado particionou as amostras em 50% para fins de treinamento e 50% para teste, sendo criados 10 conjuntos aleatórios não balanceados para fim de avaliação.

Foram criadas duas configurações: CONF1 e CONF2, que são diferenciadas pela configuração dos intervalos de valores mínimo e máximo de unidades ocultas no processo de aprendizado não-supervisionado de características das RBMs. No caso CONF1, temos um intervalo configurado com  $n \in [5, 10]$ , representando uma quantidade mínima entre 5 e 10 de características geradas pelo processo de aprendizado. Já CONF2 foi parametrizada com  $n \in [30, 45]$ , representando algo em torno de 60% à 93% das características originais. Essa configuração foi criada com o objetivo de avaliar como a quantidade de unidades ocultas pode impactar no processo de aprendizado de características mais relevantes. É importante destacar que as faixas de valores foram definidas empiricamente. Já os demais parâmetros foram configurados dentro dos intervalos:  $\eta \in [0.1, 0.9]$ ,  $\alpha \in [0.1, 0.9]$  e  $\lambda \in [10^{-5}, 10^{-2}]$ .

### 3.3 PROTOCOLO DOS EXPERIMENTOS: EXP2

Nos experimentos denominados de EXP2, as bases: LINGSPAM e CSDMC, passaram por um pré-processamento para a estratificação de um dicionário próprio baseado na seleção de características usando a técnica de Distribuição  $X^2$  (qui-quadrado). Esta foi usada para selecionar as características que, dentro da técnica, apresentavam os maiores valores de relevância dentro da distribuição. Como a SPAMBASE fica limitada a um conjunto de palavras já pré-processada em seu dicionário, essa técnica não pode ser aplicada. Após o processo de estratificação *holdout* a base LINGSPAM apresentou um dicionário de 154 palavras, ou seja 154 características, sendo estas caracterizadas pelas 154 palavras que atingiram os melhores índices de  $\chi^2(w)$ . Já a CSDMC apresenta um dicionário de 165 palavras, ou seja 165 características, sendo caracterizadas pelas 165 palavras que atingiram os melhores índices de  $\chi^2(w)$ . Diferentemente dos experimentos EXP1, outro princípio aplicado foi o balanceamento das amostras que compõem as duas classes nos conjuntos de treinamento e teste. Sendo o processo *holdout* aplicado particionando as amostras em 75% para fins de treinamento e 25% para teste, da mesma forma, foram criados 10 conjuntos aleatórios para fim de avaliação.

As configurações CONF1 e CONF2 dos experimentos EXP2 seguiram as mesmas configurações de parâmetros aplicados nos experimentos EXP1, apresentado os mesmos valores de intervalos para os parâmetros:  $n$ ,  $\eta$ ,  $\alpha$  e  $\lambda$ .

## 4 RESULTADOS EXPERIMENTAIS

São descritos os experimentos conduzidos para avaliar a robustez da abordagem proposta, bem como as análises sobre os resultados experimentais obtidos.

Inicialmente, após preparados os conjuntos de arquivos de treinamento e teste, tanto para os experimentos EXP1 quanto EXP2, cada conjunto composto por um arquivo de treinamento e teste foi denominado de RODADA, variando de 1 à 10. Estes primeiros conjuntos originais foram denominados de ORIGINAL e foram repetidos para cada uma das bases de dados: SPAMBASE, LINGSPAM e CSDMC.

Para os experimentos relacionados à EXP1, foram conduzidos experimentos a fim de aprender um conjunto mais discriminativo de características por meio das RBMs. Foram executadas três versões da RBM otimizadas com HS, IHS e PSF-HS para cada uma das bases utilizadas nos experimentos. Portanto, gerando ao final desta etapa mais três versões diferentes de cada um dos 10 conjuntos de arquivos de treinamento e teste com as características aprendidas pelas RBMs otimizadas, que foram denominadas de RBM-HS, RBM-IHS e RBM-PSF-HS, respectivamente. O mesmo procedimento foi aplicado para os experimentos relacionados à EXP2 para as bases de dados LINGSPAM e CSDMC, que antes passaram por um processo de seleção de características usando a criação de um dicionário próprio, diferentemente de EXP1, onde o mesmo dicionário pré-definido da SPAMBASE foi aplicado para todas as bases.

Os conjuntos de arquivos de treinamento e teste aprendidos pelas RBMs otimizadas, juntamente com os conjuntos originais (RBM-HS, RBM-IHS, RBM-PSF-HS e ORIGINAL), foram utilizados para alimentar o esquema tradicional de treinamento e classificação utilizando o método de classificação OPF. Em ambos os experimentos EXP1 e EXP2, foram aplicadas as configurações CONF1 e CONF2 diferenciadas pela aplicação de intervalos diferentes de neurônios ocultos  $n$ . É importante destacar que este parâmetro é responsável por determinar a quantidade de neurônios ocultos utilizado no processo de aprendizado das RBMs.

Os resultados referentes aos experimentos conduzidos sobre as bases de dados são apresentados na Seção 4.1 referentes aos dados da SPAMBASE, já a Seção 4.2 apresenta os resultados referentes a LINGSPAM e por fim na Seção 4.3 são apresentados os resultados sobre a CSDMC.

#### 4.1 RESULTADOS EXPERIMENTOS: SPAMBASE

A Tabela 3 apresenta os resultados das rodadas considerando o conjunto de dados SPAMBASE para as configurações CONF1 e CONF2 do experimento EXP1. Pelos resultados obtidos, pode-se observar que as características baseadas em RBM-HS, RBM-IHS e RBM-PSF-HS foram mais eficazes do que a ORIGINAL em todas as configurações, como observado nas RODADAS #1, #4 e #5. Entretanto, nas outras RODADAS mesmo que todas as variações RBM não tenham superado o valor ORIGINAL, pelo menos uma das técnicas baseadas nas RBMs superaram a acurácia dos dados ORIGINAL, exceto nas RODADAS #6 e #10. É importante destacar que a reconstrução das características baseada nas RBMs dos experimentos da configuração CONF1 utilizaram somente uma média de 10 características, representando 4,8 vezes menos características do que as encontradas nos dados ORIGINAL. Já nos experimentos, apenas as rodadas #6, #10 não tiveram seus resultados superados pelas variações RBMs em relação ao resultado de ORIGINAL. Embora seja observado um comportamento similar em ambos os experimentos, nota-se em CONF2 uma pequena melhora nos resultados, fator que pode ser explicado pelo número médio de características após a reconstrução ter sido maior, em média apresentando 35 características, número maior do que as 10 características encontradas em CONF1. Como o número de características da base SPAMBASE já é limitada e está pré-definida, esta não passou pelo processo de construção de um novo dicionário, não apresentando assim resultados para os experimentos EXP2. Os dados das variações RBMs que superaram os respectivos valores dos dados ORIGINAL, estão destacados em vermelho.

Tabela 3 - Resultados das configurações CONF1 e CONF2 do experimento EXP1 da SPAMBASE.

SPAMBASE (EXP1)		RBM (CONF1)			RBM (CONF2)		
RODADA	ORIGINAL	HS	IHS	PSF-HS	HS	IHS	PSF-HS
1	75,80	81,65	81,77	79,30	79,15	80,95	78,22
2	75,77	81,96	72,60	76,22	77,64	79,96	80,61
3	73,05	67,46	67,60	73,61	68,71	71,39	70,72
4	71,52	75,95	85,85	75,44	85,44	86,71	74,20
5	71,19	74,07	79,78	79,20	73,89	75,65	80,26
6	75,96	73,06	73,26	70,54	70,55	69,31	73,79
7	72,99	71,18	69,91	68,08	73,21	64,77	67,09
8	72,92	78,35	68,48	73,06	69,70	74,15	71,80
9	71,58	72,10	65,52	71,83	67,36	68,10	73,23
10	82,72	72,78	72,01	73,51	72,82	72,95	71,92
MÉDIA	74,35	74,86	73,68	74,08	73,85	74,39	74,18

Fonte: Dados da pesquisa do autor.

## 4.2 RESULTADOS EXPERIMENTOS: LINGSPAM

A Tabela 4 retrata os resultados dos experimentos EXP1 considerando a base de dados LINGSPAM. No que diz respeito à acurácia dos experimentos, apenas as rodadas #1 e #4 de CONF1 e CONF2 tiveram valores que superaram a ORIGINAL em todas as variações RBMs. Outro ponto a destacar é que na média os resultados foram inferiores aos encontrados nos experimentos do conjunto dos dados SPAMBASE, o que aponta para a não adequação do uso de um dicionário pré-definido por outra base de dados. Apenas rodadas #1 e #4 em CONF1 e CONF2 superaram em sua totalidade os resultados ORIGINAL, bem como alguns valores nas rodadas #3 e #5, como destacado em vermelho.

Tabela 4 - Resultados das configurações CONF1 e CONF2 do experimento EXP1 da LINGSPAM.

LINGSPAM (EXP1)		RBM (CONF1)			RBM (CONF2)		
RODADAS	ORIGINAL	HS	IHS	PSF-HS	HS	IHS	PSF-HS
1	62,26	71,00	65,81	80,50	77,76	70,63	66,81
2	81,83	70,71	66,40	68,22	64,78	75,98	66,19
3	59,44	58,72	64,82	58,93	57,73	60,34	66,02
4	56,50	61,42	64,65	67,39	61,88	70,09	64,57
5	62,68	61,55	56,44	66,19	68,30	71,42	60,26
6	62,55	59,63	59,43	57,52	60,26	59,72	61,21
7	67,86	66,15	64,74	63,41	69,01	66,81	66,56
8	79,22	71,09	65,91	75,44	61,76	65,49	69,18
9	66,90	64,32	56,52	60,84	59,93	57,14	66,10
10	69,52	55,65	57,02	58,81	61,96	57,44	60,26
MÉDIA	66,88	64,02	62,17	65,73	64,34	65,51	64,72

Fonte: Dados da pesquisa do autor.

Já com a aplicação do próprio dicionário de palavras, criado após aplicada a estratificação nas mensagens da LINGSPAM, foram alcançados resultados diferentes aos encontrados com a aplicação do dicionário comum de palavras usado nos experimentos EXP1, como pode ser observado na Tabela 5. Nota-se um considerável aumento nas médias de acurácia obtidas em todos os resultados, superando resultados na casa dos 60% encontrados em EXP1, com valores acima de 80%.

Em várias rodadas, algumas variações RBMs, obtiveram resultados acima de 90%, o que pode ser considerado um bom resultado. Observando a linha referente a RODADA #6, os resultados obtidos pelos dados ORIGINAL foram muito inferiores aos obtidos pelas variações RBMs, sinalizando um caminho promissor para o uso das RBMs no aprendizado de características mais relevantes. Mesmo não superando os resultados dos dados ORIGINAL as variações RBMs obtiveram resultados superior à 90% em algumas rodadas (negrito). Outro ponto é que

Tabela 5 - Resultados das configurações CONF1 e CONF2 do experimento EXP2 da LINGSPAM.

LINGSPAM (EXP2)		RBM (CONF1)			RBM (CONF2)		
RODADAS	ORIGINAL	HS	IHS	PSF-HS	HS	IHS	PSF-HS
1	96,25	88,75	<b>92,92</b>	<b>90,41</b>	85,83	<b>94,16</b>	88,75
2	96,25	80,42	64,17	88,33	59,17	61,25	87,50
3	96,25	84,58	84,17	81,67	89,58	84,17	57,50
4	98,75	<b>93,33</b>	85,42	84,58	86,67	<b>90,00</b>	86,67
5	87,92	83,75	<b>91,25</b>	79,17	82,50	85,00	56,67
6	49,17	78,75	80,83	81,25	87,08	83,33	<b>90,00</b>
7	92,08	<b>92,50</b>	84,17	85,42	81,25	82,92	80,00
8	95,42	87,92	87,08	87,92	87,50	<b>92,92</b>	88,75
9	92,92	87,08	85,42	85,00	85,00	<b>93,93</b>	82,08
10	88,33	82,50	69,16	85,00	64,58	68,75	84,58
MÉDIA	89,33	85,96	82,46	84,88	80,92	83,64	80,25

Fonte: Dados da pesquisa do autor.

houve uma evolução nos resultados de acurácia de EXP1 [62-67%] para EXP2 [80-89%].

#### 4.3 RESULTADOS EXPERIMENTOS: CSDMC

No que diz respeito à base de dados CSDMC, de acordo com a Tabela 6, observa-se que nenhum dos experimentos com as RBMs realizados em EXP1, tanto para a CONF1 e CONF2, superaram os resultados de acurácia obtidos pelas rodadas do ORIGINAL. Fato que pode ser explicado em virtude da não adequação do dicionário baseado em outra base de dados. Contudo, o principal foco deste trabalho não é obter os melhores resultados para essas base de dados, mas melhorar as taxas de classificação usando o aprendizado de características baseado em RBMs frente às características originais. Apenas em CONF1 na rodada 2, os valores das variações RBMs superaram os resultados obtidos nos dados ORIGINAL, destacado em vermelho.

Os experimentos realizados após a estratificação do dicionário próprio de palavras da base CSDMC foram muito superiores ao alcançados com a versão do dicionário compartilhado utilizado nos experimentos EXP1, efeito também observado na LINGSPAM. Como pode ser visto na Tabela 7 as médias de resultados em todas as variações das RBMs foram superiores aos resultados ORIGINAL, atingindo resultados acima de 94%, indicados em vermelho na tabela. Estes resultados também indicam um caminho promissor para o uso das variações RBMs no aprendizado de características. Os resultados encontrados nas variações RBMs da configuração CONF1 são superiores aos encontrados na configuração CONF2, ressaltando que CONF1 apresenta 4,5 vezes menos características do que CONF2. Os resultados obtidos em CONF1 foram alcançados com aproximadamente 15 vezes menos características que as encontradas nos

Tabela 6 - Resultados das configurações CONF1 e CONF2 do experimento EXP1 da CSDMC.

CSDMC (EXP1)		RBM (CONF1)			RBM (CONF2)		
RODADAS	ORIGINAL	HS	IHS	PSF-HS	HS	IHS	PSF-HS
1	65,77	62,40	61,63	62,70	51,18	53,92	60,98
2	65,26	67,64	65,36	68,23	51,89	62,42	62,73
3	63,41	57,72	58,32	57,18	53,76	45,11	60,65
4	69,65	63,68	61,03	58,86	47,75	55,56	61,68
5	66,66	62,71	62,89	62,08	43,16	46,04	64,04
6	70,45	65,43	61,25	64,55	56,87	50,58	65,75
7	61,40	61,17	60,63	60,21	45,83	52,81	62,61
8	68,43	61,85	62,63	61,08	47,88	52,88	62,12
9	65,05	63,87	61,29	60,75	54,34	58,69	60,55
10	65,32	58,22	58,69	60,32	46,41	51,43	58,53
MÉDIA	66,14	62,47	61,37	61,60	49,91	52,94	61,96

Fonte: Dados da pesquisa do autor.

dados ORIGINAL, já CONF2 com aproximadamente 3 vezes menos. Na rodada 7 não houve nenhuma das variações RBMs superando os dados ORIGINAL.

Tabela 7 - Resultados das configurações CONF1 e CONF2 do experimento EXP2 da CSDMC.

CSDMC (EXP2)		RBM (CONF1)			RBM (CONF2)		
RODADAS	ORIGINAL	HS	IHS	PSF-HS	HS	IHS	PSF-HS
1	94,53	94,08	96,71	92,01	95,41	91,86	93,05
2	75,30	96,01	96,15	95,71	96,30	95,12	95,56
3	84,17	96,45	95,56	95,41	95,86	95,27	95,71
4	95,71	94,52	96,89	96,75	94,67	95,71	95,56
5	85,50	96,60	95,12	96,75	94,67	94,53	95,86
6	91,40	95,27	96,01	95,12	93,05	93,34	92,01
7	97,04	94,82	95,27	93,49	96,01	94,08	94,67
8	78,99	94,82	95,71	94,67	94,82	95,56	94,38
9	80,77	93,49	94,23	96,01	92,75	92,90	91,72
10	94,08	93,93	94,53	93,49	92,90	93,93	93,34
MÉDIA	87,75	95,00	95,62	94,94	94,64	94,23	94,19

Fonte: Dados da pesquisa do autor.

#### 4.4 CONSIDERAÇÕES FINAIS

Como observado nos resultados apresentados na Tabela 8, mesmo quando as variações RBMs não superaram os resultados ORIGINAL, os valores ficaram bem próximos, mas é importante apontar que com menos características. Já os resultados encontrados em EXP2 da CSDMC foram superiores aos ORIGINAL e ficaram próximos à 95% (como pode ser visto nos valores destacados em vermelho), com uma proporção bem menor de características, indicando



Tabela 8 - Resultados consolidados dos experimentos EXP1 e EXP2.

-		RBM (CONF1)			RBM (CONF2)		
BASES	ORIGINAL	HS	IHS	PSF-HS	HS	IHS	PSF-HS
SPAMBASE EXP1	74,35	74,86	73,68	74,08	73,85	74,39	74,18
LINGSPAM EXP1	66,88	64,02	62,17	65,73	64,34	65,51	64,72
LINGSPAM EXP2	89,33	85,96	82,46	84,88	80,92	83,64	80,25
CSDMC EXP1	66,14	62,47	61,37	61,60	49,91	52,94	61,96
CSDMC EXP2	87,75	95,00	95,62	94,94	94,64	94,23	94,19

Fonte: Dados da pesquisa do autor.

assim, que um bom trabalho de seleção de características pode impactar resultados similares aos originais com uma proporção bem menor de características. Importante ressaltar que com menos características menor será o custo computacional e o tempo de processamento do processo de aprendizado.

Em relação aos tempos médios relacionados as fases de treinamentos e testes da aplicação do classificador OPF, pode-se observar reduções significativas principalmente na etapa do treinamento nas variações RBMs, como visto nas Tabelas 9 e 10, com os tempos médios dos experimentos EXP1 e EXP2 dentro de suas duas configurações CONF1 e CONF1, respectivamente. Ao lado dos tempos das variações RBMs estão os percentuais indicativos da redução do tempo (quando negativos), ou ao aumento do tempo (quando positivos), calculado em relação a linha dos dados ORIGINAL (ORIG.) da respectiva base. Também é importante ressaltar que os tempos de treinamento e teste são os tempos médios entre as rodadas, apenas no caso das variações RBMs, também foi aplicada a média de tempos entre as variações HS, IHS e PSF-HS para cada conjunto de bases, a fim de possibilitar a comparação entre os tempos dos dados ORIGINAL e as variações RBMs.

Tabela 9 - Resultados das medições de tempos médios de treinamento e teste dos experimentos EXP1.

EXPERIMENTOS EXP1								
BASE	CONF1				CONF1			
	TREINO		TESTE		TREINO		TESTE	
SPAMBASE (ORIG.)	0,941		0,505		0,941		0,505	
SPAMBASE (RBMs)	0,647	-31,2%	0,434	-14,1%	0,413	-56,1%	0,264	-47,7%
LINGSPAM (ORIG.)	0,382		0,198		0,382		0,198	
LINGSPAM (RBMs)	0,270	-29,3%	0,158	-20,2%	0,346	-9,4%	0,205	+3,5%
CSDMC (ORIG.)	0,775		0,441		0,775		0,441	
CSDMC (RBMs)	0,510	-34,2%	0,345	-21,8%	0,256	-67,0%	0,087	-80,2%

Tempos médios em segundos para cada configuração.

Fonte: Dados da pesquisa do autor.

Como observado nos resultados apresentados na Tabela 9 referente aos tempos de treinamento (treino) e teste dos experimentos EXP1, pode-se notar que para todas as bases houve uma redução no tempo de treinamento, mesmo no tempo relacionado a etapa de teste da CONF2 das variações RBMs da base LINGSPAM, com um percentual de 3,5% (valor destacado em vermelho na tabela) maior relacionado ao tempo dos dados ORIGINAL.

Tabela 10 - Resultados das medições de tempos médios de treinamento e teste dos experimentos EXP2.

EXPERIMENTOS EXP2								
BASE	CONF1				CONF2			
	TREINO		TESTE		TREINO		TESTE	
LINGSPAM (ORIG.)	0,117		0,022		0,117		0,022	
LINGSPAM (RBMs)	0,045	-61,5%	0,009	-59,1%	0,065	-44,4%	0,013	-40,9%
CSDMC (ORIG.)	0,942		0,177		0,942		0,177	
CSDMC (RBMs)	0,251	<b>-73,4%</b>	0,043	<b>-75,7%</b>	0,598	-36,5%	0,107	-39,5%

Tempos médios em segundos para cada configuração.

Fonte: Dados da pesquisa do autor.

Nos resultados apresentados na Tabela 10, referente aos experimentos EXP2 aplicado as bases LINGSPAM e CSDMC, foram obtidos em todas as etapas de treinamento e teste das variações RBMs tempos menores que os dados ORIGINAL. No caso, a CSDMC obteve os melhores índices de tempos, em média 73,4% menores nas variações RBMs relacionado aos dados ORIGINAL na etapa de treinamento, e 75,7% menores na etapa de teste (números destacado em negrito na tabela). Outro ponto a ser observado nestes dados está relacionado a proporção da diminuição dos tempos ser menor em CONF2 frente a CONF1, isto devido a CONF2 apresentar um número de características mais próximo ao número dos de características dos dados ORIGINAL. Mas, mesmo CONF1 tendo um número menor de características, os valores de acurácia ficaram bem próximos aos encontrados em CONF2, além de terem ficado similares aos resultados obtido com os dados ORIGINAL.

No geral, um fato importante que deve ser sinalizado é que o número características nos experimentos EXP2 foi 15 vezes menor em CONF1, e 5 vezes menor em CONF2, em relação as características originais. Os mesmos índices são mantidos na CSDMC que obteve resultados superiores aos encontrados nos dados ORIGINAL. E como observado nas relações de tempo, fica evidente que o processo de treinamento nas variações RBMs são inferiores, em alguns casos sendo perto de 3 vezes menores, fato que pode impactar positivamente também em relação ao custo computacional relacionado a esta etapa.

## 5 CONCLUSÕES

Com a realização deste trabalho, foi possível adquirir um maior conhecimento sobre os assuntos relacionados à proposta apresentada. Observou-se, ainda, que existem poucos trabalhos relacionados ao aprendizado de características utilizando arquitetura de redes com o uso de RBMs na área de segurança de rede de computadores. Neste trabalho, foram estudadas as influências do aprendizado de características mais compactas e discriminativas usando RBMs otimizadas com técnicas baseadas em HS. Doravante, são apresentadas as principais considerações e discussões acerca dos experimentos conduzidos.

Os experimentos foram conduzidos no contexto de detecção de *spam* em mensagens eletrônicas em três bases de dados públicas bem conhecidas da área: SPAMBASE, LINGSPAM e CSDMC. Em aplicações de processamento de linguagem natural, palavras são modeladas na forma de matrizes de distribuição de frequência, onde cada dimensão desta matriz representa uma característica de uma base de dados. O tamanho deste vocabulário pode ser facilmente composto por um grande número de palavras, sendo que a dimensionalidade impacta sobre o tempo de treinamento e os recursos computacionais necessários para realizar essa atividade. Por isso, a importância do uso de ferramentas com RBMs para aprender características.

Com o objetivo de avaliar os métodos propostos neste trabalho, foram conduzidos dois experimentos (EXP1 e EXP2), sendo que cada um deles apresenta duas configurações diferentes de faixas de neurônios ocultos (CONF1 e CONF2), variável que impacta diretamente no processo da quantidade de características que são aprendidas pelas variações RBMs. Para avaliar e comparar os resultados obtidos nos primeiros experimentos (EXP1), houve a aplicação do mesmo dicionário de palavras para todas as bases, no caso o dicionário definido na SPAMBASE. Já no segundo experimento (EXP2), para as bases LINGSPAM e CSDMC foi realizado um processo de seleção de características para determinar um dicionário próprio para ambas, possibilitando comparar o impacto do uso de um dicionário padrão ou dicionário próprio, no caso das bases LINGSPAM e CSDMC.

Nos resultados dos experimentos relacionados à *spam*, observa-se que podem ser alcançados resultados adequados empregando até 15 vezes menos características do que a base ORIGINAL. O uso do mesmo dicionário da SPAMBASE pode não ter sido uma boa escolha para as bases de dados LINGSPAM e CSDMC, pois mostraram resultados piores na questão da acurácia do que os encontrados na SPAMBASE em seus dados originais. Isso evidencia a importância

da aplicação de um processo prévio de seleção de características, criando assim um dicionário próprio para cada base a ser trabalhada.

O aumento da faixa de valores relacionados ao número de neurônios ocultos não impactou em melhores resultados nestes experimentos. Pôde-se observar resultados muito promissores, uma vez que existem situações em que o aprendizado de novas características pode melhorar a acurácia frente as características originais com uma vantagem considerável.

Fica evidente pelos resultados que apenas as variações RBMs nos experimentos EXP2 da CSDMC superaram em sua totalidade as acurácias de todos os conjuntos de treinamento e teste dos dados originais das bases, mas nos outros experimentos e configurações nem todas as variações RBMs aplicadas aos conjuntos superam os dados originais.

Dentro dos experimentos realizados, percebe-se que a realização de um trabalho de pré-seleção de características mais significativas pode trazer melhores resultados. Neste caso, é importante em trabalhos futuros uma maior investigação sobre este comportamento. Outro ponto importante a destacar é que uma variação RBM baseada na otimização HS pode ser mais adequada a um conjunto de dados, mas não a outro, e vice-versa.

Diante das análises e considerações apresentadas, pode-se finalizar o estudo indicando que os resultados apresentados apontam para um futuro promissor no uso de variações RBMs no emprego do aprendizado não-supervisionado de características, e que avanços de pesquisas neste sentido devem ocorrer.

## 5.1 PRINCIPAIS CONTRIBUIÇÕES

Ao final desta dissertação, as seguintes contribuições podem ser elencadas:

- Avaliação da robustez das RBMs na tarefa de aprendizado de características em modo não-supervisionado;
- Aplicação do refinamento dos parâmetros das RBMs por meio da aplicação de meta-heurísticas baseadas em algoritmos de otimização baseados em técnicas HS;
- Diminuição do tempo da fase treinamento e do custo computacional relacionado a este processo, através da aplicação do aprendizado não-supervisionado de características;
- Aplicação do aprendizado não-supervisionado de características para detecção de conteúdo malicioso, especificamente na área de segurança de redes de computadores.

## 5.2 TRABALHOS FUTUROS

Dentre as sugestões de trabalhos futuros, podem-se destacar:

- Aplicar as técnicas de aprendizado de características em modo não-supervisionadas usando RBMs com mais camadas intermediárias, usando assim, suas variações em redes profundas DBN;
- Realizar experimentos com intervalos diferentes para o número de unidades escondidas para analisar o impacto desta variável na eficiência do processo de aprendizado não-supervisionado de características, bem como explorar variações nos outros parâmetros apresentados;
- Empregar outros classificadores afim de comparar resultados relacionados ao acurácia e ao tempo do processo de treinamento;
- Testar outros algoritmos de otimização para refinamento dos parâmetros das RBMs;
- Utilizar o método estudado e aplicado para o aprendizado de características em outros tipos de anomalias em redes de computadores afim de avaliar os resultados e comparar com outros métodos;
- Aplicar a metodologia proposta em outras bases de dados de conteúdo malicioso para avaliar se os mesmos comportamentos são observados em relação aos experimentos constantes deste trabalho.

## 6 PUBLICAÇÕES REALIZADAS

Cabe ressaltar que, diante dos trabalhos realizados durante o programa de mestrado, foi submetido um artigo relacionado ao tema da proposta intitulado “*Parameter-setting Free Harmony Search Optimization of Restricted Boltzmann Machines and its Applications to Spam Detection*”, o qual foi apresentado na conferência: “*12th International Conference Applied Computing 2015*”, Dublin, Irlanda. O referido artigo foi agraciado com o prêmio de “*Best Paper Award*”. Mediante ao prêmio recebido, houve o convite para estender o trabalho no periódico internacional “*IADIS International Journal On Computer Science And Information Systems*”, ISSN: 1646-3692, sendo a extensão do artigo submetido sob o título “*Learning Spam Features Using Restricted Boltzmann Machines*”, ambos servido de base para a realização deste trabalho.

A seguir são listados os artigos publicados que tiveram a participação do autor e estão relacionados ao tema central deste trabalho:

- Artigo publicado em Conferência Internacional:
  - SILVA, L. A.; COSTA, K. A. P.; RIBEIRO, P. B.; ROSA, G.; PAPA, J. P. “*Parameter-setting Free Harmony Search Optimization of Restricted Boltzmann Machines and Its Applications to Spam Detection*”. 12th International Conference on Applied Computing, 2015, Dublin, Irlanda. p. 143-150.
- Artigos publicados em Periódicos Internacionais:
  - SILVA, L. A.; COSTA, K. A. P.; RIBEIRO, P. B.; ROSA, G.; PAPA, J. P. “*Learning Spam Features Using Restricted Boltzmann Machines*”. IADIS International Journal on Computer Science and Information Systems, 2016. Vol. 11, No. 1, p. 99-114. ISSN: 1646-3692.
  - SILVA, L. A.; COSTA, K. A. P.; RIBEIRO, P. B. ; FERNANDES, D.; PAPA, J. P. “*On the Feasibility of Optimum-Path Forest in the Context of Internet-of-Things-based Applications*”. Recent Progress in Space Technology, 2015.
  - COSTA, K.; SILVA, L.; MARTINS, G.; ROSA, G.;PIRES, R.; PAPA, J. “*On the Evaluation of Restricted Boltzmann Machines for Malware Identification*”. International Journal Of Information Security Science. p. 69-81.
- Participação em artigos em Conferência, Simpósios e Congresso:

- COSTA, K. A. P. ; SILVA, L. A. ; MARTINS, G. B. ; ROSA, G. H. ; PEREIRA, C. R. ; PAPA, J. P. “*Malware Detection in Android-based Mobile Environments using Optimum-Path Forest*”. 14th International Conference on Machine Learning and Applications, 2015, Miami, FL, USA.
- RIBEIRO, P. B.; PASSOS, L. A.; SILVA, L. A.; COSTA, K. A. P.; PAPA, J. P. ; ROMERO, R. A. F. “*Unsupervised Breast Masses Classification through Optimum-Path Forest*”. IEEE 28th International Symposium on Computer Based Medical Systems (CBMS), 2015, São Carlos, SP, Brasil. p. 238-243.

## REFERÊNCIAS

- ACKLEY, D.; HINTON, G.; SEJNOWSKI, T. J. A learning algorithm for boltzmann machines. *Connectionist Models and Their Implications: Readings from Cognitive Science*, p. 285–307, 1988.
- ALLEN, J.; CHRISTIE, A.; FITHEN, W.; MCHUGH, J.; PICKEL, J.; ELLIS, J.; HAYES, E.; MARELLA, J.; WILLKE, B. State of the practice of intrusion detection technologies. *Carnegie Mellon Software Engineering Institute*, 2000.
- ALMEIDA, T. A.; YAMAKAMI, A. Advances in spam filtering techniques. In: \_\_\_\_\_. *Computational Intelligence for Privacy and Security*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. p. 199–214.
- ALMEIDA, T. A.; YAMAKAMI, A.; ALMEIDA, J. Evaluation of approaches for dimensionality reduction applied with naive bayes anti-spam filters. In: *International Conference on Machine Learning and Applications, 2009 (ICMLA)*. Miami, FL, USA: [s.n.], 2009. p. 517–522.
- ANDROUTSOPOULOS, I.; KOUTSIAS, J.; CHANDRINOS, K.; PALIOURAS, G.; SPYROPOULOS, C. An evaluation of naive bayesian anti-spam filtering. *Proceedings of the workshop on Machine Learning in the New Information Age*, 2000. Disponível em: <[citeseer.ist.psu.edu/androustopoulos00evaluation.html](http://citeseer.ist.psu.edu/androustopoulos00evaluation.html)>.
- ASSIS, J. M. D. C. *DETECÇÃO DE E-MAILS SPAM UTILIZANDO REDES NEURAIAS ARTIFICIAIS*. Itajubá; Brasil: Mestre em Ciências em Engenharia Elétrica. UNIVERSIDADE FEDERAL DE ITAJUBÁ, 2006.
- CARREIRA-PERPINAN, M. A.; HINTON, G. E. On contrastive divergence learning. In: COWELL, R. G.; GHAHRAMANI, Z. (Ed.). *10th International Workshop on Artificial Intelligence and Statistics*. [S.l.: s.n.], 2005. p. 33–40.
- DAHL, G. E.; ADAMS, R. P.; LAROCHELLE, H. Training restricted Boltzmann machines on word observations. *Proceedings of the 29th International Conference on Machine Learning (ICML- 2012)*, Omnipress, New York, NY, USA, p. 679–686, July 2012.
- DAS KANISHKA BHADURI, P. V. K. Distributed anomaly detection using 1-class svm for vertically partitioned data. *Statistical Analysis and Data Mining*, John Wiley and Sons, Inc., New York, NY, 2011.
- DASH, M.; LIU, H. Feature selection for classification. *Intelligent Data Analysis*, v. 1, p. 131–156, 1997.
- FALCÃO, A.; STOLFI, J.; LOTUFO, R. The image foresting transform: Theory, algorithms, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 19–29,



2004.

FIORE, U.; PALMIERI, F.; CASTIGLIONE, A.; SANTIS, A. D. Network anomaly detection with the restricted boltzmann machine. *Neurocomputing: Elsevier Science Publishers*, Elsevier Science Publishers B. V., p. 13–23, 2013.

FISCHER, A.; IGEL, C. Training restricted boltzmann machines: An introduction. *Pattern Recognition*, v. 47, n. 1, p. 25–39, 2014.

GEEM, Z. W.; KIM, J. H.; LOGANATHAN, G. A new heuristic optimization algorithm: Harmony search. *SIMULATION*, v. 76, n. 2, p. 60–68, 2001. Disponível em: <<http://sim.sagepub.com/content/76/2/60.abstract>>.

GEEM, Z. W.; SIM, K.-B. Parameter-setting-free harmony search algorithm. *Applied Mathematics and Computation*, v. 217, n. 8, p. 3881 – 3889, 2010. ISSN 0096-3003.

GROUP, C. *Spam corpus: Spam email datasets*. 2010. Disponível em: <<http://csmining.org/index.php/spam-email-datasets-.html>>.

HAJINOROOZI, M.; JUNG, T.; LIN, C.; HUANG, Y. Feature extraction with deep belief networks for driver's cognitive states prediction from eeg data. *IEEE China Summit and International Conference on Signal and Information Processing*, p. 812–815, 2015.

HE, S.; WANG, S.; LAN, W.; FU, H.; JI, Q. Facial expression recognition using deep boltzmann machine from thermal infrared images. *Humaine Association Conference on Affective Computing and Intelligent Interaction*, p. 239–244, 2013.

HINTON, G. Training products of experts by minimizing contrastive divergence. *Neural Computation*, MIT Press, Cambridge, MA, USA, v. 14, n. 8, p. 1771–1800, 2002. ISSN 0899-7667.

HINTON, G. A practical guide to training restricted boltzmann machines. *Neural Networks: Tricks of the Trade*, p. 599–619, 2012.

HINTON, G.; SALAKHUTDINOV, R. R. Reducing the dimensionality of data with neural networks. *Journal Science*, v. 313, n. 5786, p. 504–507, 2006.

HINTON, G. E. Training products of experts by minimizing contrastive divergence. *Neural Computation*, v. 14, n. 8, p. 1771–1800, 2002.

HINTON, G. E.; OSINDERO, S.; TEH, Y.-W. A fast learning algorithm for deep belief nets. *Neural Computation*, v. 18, n. 7, p. 1527–1554, 2006.

HIRA, Z. M.; GILLIES, D. F. A review of feature selection and feature extraction methods applied on microarray data. *Adv. Bioinformatics*, v. 2015, p. 198363:1–198363:13, 2015.

IDC. *IDC Threat Intelligence*. [S.l.], 2012.

JATANA, N.; SHARMA, K. Bayesian Spam Classification: Time Efficient Radix Encoded Fragmented Database Approach. *International Conference on Computing for Sustainable Global Development (INDIACom)*, p. 939–942, 2014.

JOACHIMS, T. Text categorization with support vector machines: Learning with many relevant features. In: *Proceedings of the 10th European Conference on Machine Learning*. London, UK, UK: Springer-Verlag, 1998. (ECML '98), p. 137–142. ISBN 3-540-64417-2.

KAKADE, A.; KHARAT, P.; GUPTA, A.; BATRA, T. Spam filtering techniques and MapReduce with SVM: A study. *Asia-Pacific Conference on Computer Aided System Engineering (APCASE)*, p. 59–64, 2014.

LAROCHELLE, H.; BENGIO, Y. Classification using discriminative restricted boltzmann machines. *Proceedings of the 25th International Conference on Machine Learning (ICML 2008)*, ACM, New York, NY, USA, p. 536–543, 2008.

LAROCHELLE, H.; MANDEL, M.; PASCANU, R.; BENGIO, Y. Learning algorithms for the classification restricted boltzmann machine. *The Journal of Machine Learning Research*, JMLR.org, v. 13, n. 1, p. 643–669, 2012.

LICHMAN, M. *UCI Machine Learning Repository*. 2013. Disponível em: <<http://archive.ics.uci.edu/ml>>.

LOPES, N.; RIBEIRO, B.; GONCALVES, J. Restricted boltzmann machines and deep belief networks on multi-core processors. *International Joint Conference on Neural Networks*, p. 1–7, 2012.

MAHDAVI, M.; FESANGHARY, M.; DAMANGIR, E. An improved harmony search algorithm for solving optimization problems. *Applied Mathematics and Computation*, v. 188, n. 2, p. 1567 – 1579, 2007.

MOHAMAD, M.; SELAMAT, A. An evaluation on the efficiency of hybrid feature selection in spam email classification. In: *Computer, Communications, and Control Technology (I4CT), 2015 International Conference on*. [S.l.: s.n.], 2015. p. 227–231.

PAN, G.; QIAO, J.; CHAI, W.; DIMOPOULOS, N. An improved rbm based on bayesian regularization. *International Joint Conference on Neural Networks*, p. 2935–2939, 2014.

PAPA, J. P.; FALCÃO, A. X.; ALBUQUERQUE, V. H. C.; TAVARES, J. M. R. S. Efficient supervised optimum-path forest classification for large datasets. *Pattern Recognition*, Elsevier Science Inc., New York, NY, USA, v. 45, n. 1, p. 512–520, 2012.

PAPA, J. P.; FALCÃO, A. X.; SUZUKI, C. T. N. Supervised pattern classification based on optimum-path forest. *International Journal of Imaging Systems and Technology*, John Wiley & Sons, Inc., New York, NY, USA, v. 19, n. 2, p. 120–131, 2009. ISSN 0899-9457.

PAPA, J. P.; ROSA, G. H.; COSTA, K. A. P.; MARANA, A. N.; SCHEIRER, W.; COX, D. D. On the model selection of bernoulli restricted boltzmann machines through harmony search. *Proceedings of the Genetic and Evolutionary Computation Conference*, p. 1449–1450, 2015.

PAPA, J. P.; ROSA, G. H.; MARANA, A. N.; SCHEIRER, W.; COX, D. D. Model selection for discriminative restricted boltzmann machines through meta-heuristic techniques. *Journal of Computational Science*, v. 9, p. 14–18, 2015.

PAPA, J. P.; SCHEIRER, W.; COX, D. D. Fine-tuning deep belief networks using harmony

search. *Elsevier Science Inc.: Applied Soft Computing*, 2015. ISSN 1568-4946.

ROCHA, L.; CAPPABIANCO, F.; FALCÃO, A. X. Data Clustering as an Optimum-Path Forest Problem with Applications in Image Analysis . *Wiley Periodicals, Inc.*, v. 19, p. 50–68, 2009.

RODRIGUES, D. *Seleção de Características Utilizando Algoritmos Evolucionistas e suas Aplicações em Reconhecimento de Padrões*. [S.l.]: Universidade Estadual Paulista Júlio de Mesquita Filho, 2014.

SARIKAYA, R.; HINTON, G. E.; DEORAS, A. Application of deep belief networks for natural language understanding. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, IEEE Press, Piscataway, NJ, USA, v. 22, n. 4, p. 778–784, 2014.

SEBASTIANI, F. Machine learning in automated text categorization. *ACM Comput. Surv.*, ACM, New York, NY, USA, v. 34, n. 1, p. 1–47, 2002.

SILVA, L. A.; COSTA, K. A. P.; RIBEIRO, P. B.; ROSA, G. H.; PAPA, J. P. Parameter-setting free harmony search optimization of restricted boltzmann machines and its applications to spam detection. *12th International Conference Applied Computing*, p. 142–150, 2015.

SILVA, L. A.; COSTA, K. A. P.; RIBEIRO, P. B.; ROSA, G. H.; PAPA, J. P. Learning spam features using restricted boltzmann machines. *IADIS International Journal on Computer Science and Information Systems*, v. 11, n. 1, p. 99–114, 2016.

TIELEMAN, T. Training restricted boltzmann machines using approximations to the likelihood gradient. In: *25th International Conference on Machine Learning*. New York, USA: ACM, 2008. (ICML '08), p. 1064–1071.

TIELEMAN, T.; HINTON, G. E. Using fast weights to improve persistent contrastive divergence. In: *26th Annual International Conference on Machine Learning*. New York, USA: ACM, 2009. (ICML '09), p. 1033–1040.

WELLING, M.; ROSEN-ZVI, M.; HINTON, G. Exponential family harmoniums with an application to information retrieval. *Advances in Neural Information Processing Systems 17*, MIT Press, p. 1481–1488, 2005.

WHITMAN, M. E.; MATTORD, H. J. *Principles of Information Security*. Boston, MA: Course Technology Press, 2004. ISBN 0619216255.

YANG, J.; PENG, L.; LIU, T. Anti-spam Model Based on AIS in Cloud Computing Environments. *Computer modelling and new technologies*, v. 18, p. 97–102, 2014.

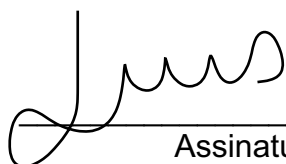
YU, C.; JIAN, Z.; BO, Y.; DEYUN, C. A novel principal component analysis neural network algorithm for fingerprint recognition in online examination system. In: *Information Processing, 2009. APCIP 2009. Asia-Pacific Conference on*. [S.l.: s.n.], 2009. v. 1, p. 182–186.

ZHOU, S.; CHEN, Q.; WANG, X. Discriminative deep belief networks for image classification. *17th IEEE International Conference on Image Processing*, p. 1561–1564, 2010.

## TERMO DE REPRODUÇÃO XEROGRÁFICA

Autorizo a reprodução xerográfica do presente Trabalho de Conclusão, na íntegra ou em partes, para fins de pesquisa.

São José do Rio Preto, 16 / 11 / 2016



Assinatura do autor