Asociación Argentina
de Mecánica Computacional

Mecánica Computacional Vol XXIX, págs. 9521-9540 (artículo completo)
Eduardo Dvorkin, Marcela Goldschmit, Mario Storti (Eds.)
Buenos Aires, Argentina, 15-18 Noviembre 2010

TÉCNICA DE REGRESSÃO LINEAR MÚLTIPLA E MÉTODOS ESTATÍSTICOS DE SELEÇÃO DE CARACTERÍSTICAS NA REDUÇÃO DE DIMENSIONALIDADE NA CLASSIFICAÇÃO DE E-MAILS

Alisson M. Silva[a,b], **Gray F. Moita**[a] **and Paulo E. M. Almeira**[a]

[a]*LSI - Laboratório de Sistemas Inteligentes, CEFET-MG - Centro Federal de Educação Tecnológica de Minas Gerais, Av. Amazonas, 7675, Nova Gameleira, 30.510-000, Belo Horizonte, MG, Brasil, alisson@lsi.cefetmg.br, gray@lsi.cefetmg.br, pema@lsi.cefetmg.br, http://www.lsi.cefetmg.br*

[b]*Departamento de Ciências Exatas, IFMG - Instituto Federal de Minas Gerais - Campus Bambuí, Faz. Varginha - Rodovia Bambuí/Medeiros - Km 05 - 38900-000 - Bambuí - MG - Brasil, alisson.marques@ifmg.edu.br http://www.cefetbambui.edu.br*

Palavras-chave: StepWise, E-mail, Spam, Redes Neurais Artificiais.

Resumo. Este trabalho apresenta a análise do uso da técnica de regressão linear múltipla *StepWise* empregada em conjunto com métodos estatísticos de seleção características em um modelo neural de filtro anti-*spam*. O objetivo do uso da técnica de regressão linear múltipla é diminuir o número de características empregadas no agente classificador, reduzindo o custo computacional e melhorando o desempenho. Os métodos de seleção de características empregados foram: Informação Mútua, *QUI statistic*, e variações do método Distribuição por Frequência. Como agente classificador utilizou-se as redes neurais *MultiLayer Perceptron*. Os resultados dos experimentos realizados mostraram que a técnica *StepWise* conseguiu alcançar o objetivo proposto, reduzir o custo computacional e melhorar o desempenho do classificador.

1 INTRODUÇÃO

Cada vez mais presente na vida das pessoas, a internet revoluciona a maneira de se obter informações, de se fazer negócios e, até mesmo, a de se relacionar. Vários serviços são disponibilizados pela internet, entre eles o *e-mail* ou correio eletrônico, que pode ser definido como uma forma de criar, enviar e receber mensagens por intermédio de sistemas eletrônicos. O correio eletrônico foi criado antes da internet, mas seu uso generalizado se deu com o surgimento e popularização desta, tornando-se um dos mais importantes meios de comunicação.

A popularização do correio eletrônico fez com que esse serviço se tornasse bastante utilizado para envio de *spam* - termo utilizado para denominar mensagens eletrônicas que, na maioria das vezes, são de intuito publicitário, visando a promoção de serviços, produtos ou eventos, e que são enviadas sem o consentimento prévio dos destinatários. Cormack and Lynam (2005) definem o *spam* como *e-mail* não solicitado, emitido de forma indiscriminada, direta ou indiretamente, por um remetente que não tem nenhum relacionamento com o destinatário. O *spam* pode ser considerado o equivalente eletrônico das correspondências indesejadas e dos telefonemas de *telemarketing* não solicitados. Para Cranor and LaMacchia (1998), os principais fatores que contribuem para o crescimento do número de *spam* são a facilidade de enviá-lo para um grande número de destinatários e de se obter endereços de *e-mails* válidos, além do baixo custo de envio.

Estudos realizados pela IronPort (2008) indicam que mais de 55% dos usuários de *e-mail* dizem ter perdido a confiança no *e-mail* por causa do *spam*. De acordo com esse mesmo estudo, as tendências dos *spams* podem ser caracterizadas por um aumento no número e tipo dos alvos e, também, pela sofisticação dos ataques. A pesquisa informa ainda que o volume diário de *spams* é superior a 120 bilhões, significando aproximadamente 20 mensagens de *spam* por dia para cada pessoa no planeta. Além disso, o *spam* tornou-se mais perigoso, isto é, se no passado eles apenas vendiam algum tipo de produto, em 2007 mais de 83% dos *spams* continham URL. A ocorrência de vírus e *malware* baseados em URL aumentaram 256%.

Conforme visto, os *spams* podem causar prejuízos às pessoas e às empresas. De acordo com estudo estatístico publicado por Evett (2006), 40% dos *e-mails* em 2006 eram *spam* e a estimativa para 2007 era de 63%, sendo que essa previsão foi superada em fevereiro de 2007. A pesquisa da Marshal (2007) revela que 85% dos *e-mails* trafegados no mês em questão eram *spam*. Em 2002 o custo para usuários não corporativos foi de U$ 225 milhões e para usuários corporativos de U$ 8.9 bilhões. Outro dado interessante nessa pesquisa é que 16% dos usuários trocam de *e-mails* devido ao *spam* e 8% compram por meio do *spam*.

Estudo realizado em julho de 2003 pela Nucleus (2003) intitulado *Spam: The Silent ROI Killer* (*Spam*: o assassino silencioso do retorno de investimento) com 76 empresas norte-americanas sobre o impacto do *spam* no trabalho dos funcionários revelou que os *spams* diminuem a produtividade em 1,4%. Uma nova pesquisa realizada em maio de 2004 com as mesmas empresas revela que esse percentual subiu para 3,2% em 10 meses. O custo anual estimado do *spam* por usuário é de U$1.934. Outro estudo da Nucleus (2007) o *Spam: The Repeat Offender de 2007*, estima que as empresas americanas perdem U$70 bilhões por ano devido a queda de produtividade causada pelo *spam*.

Segundo estatísticas da Symantec (2008) divulgadas no relatório de agosto de 2008, 78% dos *e-mails* que circularam em julho de 2008 são *spams*. Em julho de 2007 esse número era de 66%. O percentual de *spam* circulando pela rede pode variar entre 30% e 90%, dependendo do dia, semana ou mês. Diferentes empresas e países também recebem percentuais variados de

spam. Nas estatísticas do *Trace Team*[1] o percentual de *spams* flutua entre 75% a 95% (Marshal, 2008).

Segundo dados da Marshal (2008), os Estados Unidos são a principal fonte de *spam* com mais de 16% do total de *spams* enviados e a Europa lidera o *ranking* entre os continentes com 36%. O Brasil é o quinto país em envio de *spam* com 5,8%. As informações sobre os países e os continentes que mais enviam *spam* podem ser vistas respectivamente nas figuras 1a e 1b.

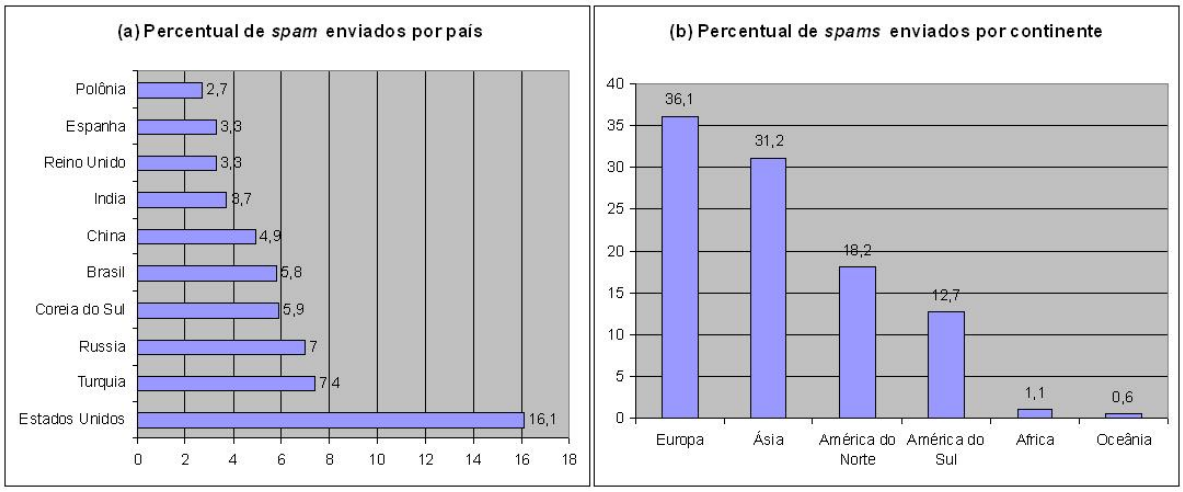

Figure 1: Percentual de *spams* enviados por país (a) e por continente (b).

Adaptado de Marshal (2008)

Os estudos apresentados, assim como os demais explorados neste trabalho, mostram os problemas e os prejuízos causados pelos *spams*, bem como a importância de se encontrar soluções que consigam minimizá-los. Segundo Ozgur et al. (2004), vários métodos para identificar e classificar os *spams* foram propostos, porém nenhum deles é completamente satisfatório.

Neste contexto, este artigo apresenta um estudo da técnica de regressão linear múltipla (*Step-Wise*) e de métodos estatísticos de seleção de características empregados na redução de dimensionalidade na classificação de *e-mails*. Como agente classificador foram utilizadas as redes neurais artificiais.

2 TÉCNICAS UTILIZADAS POR *SPAMMERS*

Os *spammers*, inicialmente, enviavam suas mensagens diretamente aos usuários, sem nenhum tipo de disfarce. Com o surgimento e evolução dos filtros anti-*spam*, esse tipo de *e-mail* era facilmente identificado e bloqueado. Em resposta à criação dos filtros, os *spammers* começaram a utilizar endereços falsos e a forjar outras informações nas mensagens.

As técnicas utilizadas para o envio de *spam* desenvolveram-se em resposta à evolução dos filtros anti-*spam*. Assim que as empresas de segurança criam filtros eficientes, os *spammers* mudam suas táticas de forma a evitar esses novos filtros. Dessa forma, ocorre um padrão circular previsível, com os remetentes de *spam* reinvestindo seus lucros no desenvolvimento de novas técnicas para burlar os novos filtros anti-*spam* (Kaspersky, 2008).

Atualmente, as principais técnicas empregadas pelos remetentes de *spam* para enganar os filtros são (Wittel and Wu, 2004):

[1]Threat Research and Content Engineering.

- **Tokenização:** o *spammer* modifica a característica da palavra, adicionando caracteres inválidos ou utilizando acentuação incorreta com o objetivo de enganar o filtro. No entanto, ele a modifica de forma que o usuário consiga entender a palavra. Por exemplo: v.ia.g ra; V I A G R A; Gânhè dinheíro fãcíl;

- **Texto invisível:** o remetente de *spam* esconde dentro do *e-mail* um texto considerado de mensagem legítima para enganar o filtro. Na visualização da mensagem no gerenciador de *e-mail* esse texto não é exibido, sendo visualizada, pelo usuário, somente o *spam*;

- **Imagens com texto:** técnica que envolve o envio das mensagens do *spam* dentro de imagens. Para detectar este tipo de *spam* é necessário adotar o método de reconhecimento óptico dos caracteres (*Optical Character Recognition - OCR*) que escanea as imagens do *e-mail* com o objetivo de identificar letras e palavras. Os *spammers* também utilizam a inserção das imagens em servidores gratuitos, referenciando-as nas mensagens em vez de anexá-las. Assim, dificultam ainda mais o trabalho dos filtros;

- ***Tags* HTML inválidos:** os *spammers* utilizam *tags* HTML inválidas para fazer com que os textos considerados legítimos sejam analisados pelo filtro, sendo exibidos no *e-mail* as mensagens de *spam*;

- **Comentários HTML:** os *spammers* escondem as palavras usando comentários HTML ou outras *tags* que serão removidas pelo gerenciador de *e-mail* na exibição da mensagem. Assim, a palavra é exibida corretamente e o filtro é enganado;

- **Mensagem bilíngue:** a mensagem é enviada em duas línguas distintas, ambas são apresentadas na leitura do *e-mail*;

- **Texto redundante:** uma mensagem legítima é enviada em texto plano e, em HTML, a mensagem *spam*. O gerenciador de correio por padrão mostra a segunda mensagem. O filtro anti-*spam* analisa as duas e o *e-mail* é considerado legítimo.

3 TÉCNICAS DE DETECÇÃO

Por meio das técnicas empregadas pelos *spammers*, percebe-se que a tarefa de identificar e bloquear os *spams* está cada dia mais complexa. Para conseguir identificar os *spams*, as técnicas de detecção focam diferentes características do *spam*, como a origem, forma, conteúdo e comportamento.

Um dos principais problemas enfrentados na detecção de *e-mails* são as classificações incorretas, *e-mails* legítimos classificados como *spam* e vice-versa. Os erros nos processos de classificação podem ser separados em dois tipos:

- **Falso Negativo:** acontece quando o filtro anti-*spam* detecta um *spam* como se fosse uma mensagem legítima, de modo que o usuário recebe em sua caixa de mensagens um *spam* que foi considerado como mensagem legítima;

- **Falso Positivo:** ocorre quando uma mensagem legítima é classificada como *spam*.

Apesar de trazer aborrecimentos, o falso negativo não é um problema grave. O usuário pode apagar a mensagem classificada incorretamente da sua caixa de entrada. Do contrário, o falso positivo pode ser um problema grave. De acordo com as configurações do filtro anti-*spam*, as

mensagens classificadas como *spam* podem ser enviadas a uma pasta específica, para posterior análise do usuário, ou até mesmo serem excluídas automaticamente após a classificação. O primeiro caso faz com que o usuário necessite verificar constantemente a pasta de mensagens *spam* em busca de mensagens legítimas. No segundo, o usuário fica sem acesso à mensagem. Dessa forma, os falsos positivos são altamente indesejáveis e devem ser evitados. Os falsos negativos, apesar de não serem tão graves, também devem ser diminuídos (Assis, 2006).

Os principais métodos utilizados para detectar os *spams* podem ser divididos em listas de bloqueio, *Greylisting*, *DomainKeys Identified Mail* e filtros de conteúdo.

3.1 Listas de Bloqueio

As listas de bloqueio contém endereços de *e-mails*, endereços IP e domínios que são conhecidos pela prática de enviar *spam*. Os servidores configurados com essas listas recusam automaticamente qualquer *e-mail* enviado por remetentes que lá estejam. As principais listas de bloqueio são:

O maior problema das listas de bloqueio é o elevado número de falsos positivos. De acordo Teixeira (2004), a implementação dessas listas, mais especificamente as negras, tem sofrido retaliações por meio de ações judiciais movidas por empresas que tiveram seus servidores incluídos nesse tipo de lista e por ataques de negação de serviço provocados por *spammers*.

3.2 *Greylisting*

Esta técnica consiste em recusar temporariamente uma mensagem e aguardar por sua retransmissão. Ao receber um *e-mail*, o MTA (*Mail Transfer Agents*) que utiliza *Greylisting* verifica o IP do servidor de origem, o *e-mail* do remetente e do destinatário. Se esse conjunto de informações verificadas não estiver na lista de permitidos, o servidor rejeita a mensagem respondendo com uma falha temporária e solicita que o *e-mail* seja reenviado dentro de um determinado tempo. Depois de certo número de reenvios com sucesso o conjunto é adicionado à lista de permitidos e deixa de ser filtrado.

O processo de *Greylisting* constitui um grande avanço no combate ao *spam*. Segundo informações disponíveis na internet, essa técnica rejeita aproximadamente 95% dos *spams*. Sua vantagem é o baixo custo computacional, sendo uma boa opção como primeira linha de defesa, antes de uma técnica com controle de conteúdo. As suas desvantagens estão na demora inicial para entrega dos *e-mails* legítimos que não fazem parte da lista de permitidos e nos servidores que não respeitam a RFC do SMTP (RFC 2821)[2], pois não conseguem enviar *e-mails* para os servidores com *Greylisting*. Para Levine (2005), um projeto de *Greylisting* implementado com cuidado pode minimizar a quantidade de mensagens legítimas não recebidas com pouca perda de efetividade no bloqueio de *spams*.

3.3 *DomainKeys Identified Mail* (DKIM)

Os *spammers* utilizam domínios e endereços de *e-mails* falsificados para o envio de *spams*. O DKIM emprega um método que consiste em assinar digitalmente as mensagens, garantindo a autenticidade do remetente.

De acordo com estudo realizado pela IronPort (2006), em 2005 apenas 1% dos *e-mails* eram autenticados empregando o DKIM. Em 2006, 9% dos *e-mails* autenticados empregavam esse método. Essa técnica usada isoladamente pode não resolver o problema dos *spams*, porém muito contribui para minimizar os problemas por eles causados. Informações complementares

[2]http://www.ietf.org/rfc/rfc2821.txt.

acerca desse tema podem ser obtidas nas RFC 4686 - *Analysis of Treats Motivating DomainKeys Idenfied Mail (DKIM)*[3] e RFC 4871 - *DomainKeys Identified Mail Signatures*[4].

3.4 Filtros de Conteúdo

Os filtros de conteúdo são aplicações capazes de separar os *e-mails* com base em seu conteúdo. São mais complexos e eficazes, pois utilizam técnicas de categorização de texto e aprendizagem de máquina baseadas em inteligência artificial para analisar o conteúdo dos *e-mails* e classificá-los. Nessa técnica, o classificador aprende, a partir de uma base de dados previamente preparada, e utiliza esse conhecimento para analisar e classificar os documentos ainda não vistos. Os principais métodos que utilizam essa técnica são: Algoritmos *Naive Bayesian* (Kosmopoulos et al., 2008; Zhang et al., 2004); SVM - *Support Vector Machine* (Rios and Zha, 2004; Zhang et al., 2004); K-NN - K-Vizinhos Mais Próximos (Andrade, 2006; Trudgian and Yang, 2004); Árvores *Boosting* (Zhang et al., 2004; Carreras and Marquez, 2001); Aprendizado Baseado em Memória Zhang et al. (2004); Sakkis et al. (2003); *Rocchio* (Drucker et al., 1999) e; Sistemas Imunológicos Artificiais (Guzella et al., 2005).

A partir dos resultados apresentados pelos diversos autores listados anteriormente, pode-se concluir que todos esses métodos são relevantes. Entretanto, a taxa de falsos positivos foi alta, na maioria deles, na faixa de 3%, o que não é bom para o usuário final. Outras técnicas podem ser empregadas para detecção de *e-mails spams*. No escopo deste estudo, serão empregadas como classificador as redes neurais artificiais (RNA).

4 CLASSIFICAÇÃO DE *E-MAILS*

A classificação é uma técnica utilizada para atribuir automaticamente um conjunto de textos a uma ou mais categorias predefinidas. A aplicação mais comum é na indexação de textos, sistemas de *data mining*, categorização de mensagens, notícias, resumos e arquivos de publicações periódicas. Rizzi et al. (2000) define a classificação de texto como uma técnica usada, principalmente, para descoberta do conhecimento, cujo objetivo é classificar documentos em relação a um conjunto de categorias predefinidas. É uma técnica para atribuir automaticamente um documento textual a um ou mais conjuntos.

O processo de classificação é menos complexo quando executado por seres humanos, devido à relativa facilidade em inferir conceitos a partir das palavras contidas nos documentos. No entanto, quando o número de documentos é grande o processo, apesar de simples, pode se tornar bastante demorado. Nos sistemas computacionais, o processo de classificação envolve técnicas para extrair as informações mais relevantes de cada categoria e utilizar estas informações para ensinar o sistema a classificar corretamente os documentos. O processo aplicado na classificação de mensagens eletrônicas, pode ser dividido em cinco etapas:

1. **Conjunto de dados e categorias:** esta etapa consiste em selecionar o conjunto de dados que será utilizado no processo de treinamento e teste do sistema e também as categorias presentes no conjunto;

2. **Preparação:** preparação ou pré-processamento é o processo de uniformização das informações presentes no conjunto de dados em que cada documento é analisado com o objetivo de remover as informações irrelevantes como acentuação, caracteres especiais, figuras, entre outros;

[3]http://www.ietf.org/rfc/rfc4686.txt.
[4]http://www.ietf.org/rfc/rfc4671.txt.

3. **Seleção das características:** visa selecionar, através de métodos estatísticos, as palavras mais relevantes, isto é, as que melhor representam as classes definidas;

4. **Vetor de características:** nessa etapa, as palavras selecionadas na etapa anterior são indexadas e utilizadas para compor o vetor de entrada para o agente classificador.

5. **Classificador:** é nesta etapa que efetivamente ocorre a classificação da mensagem.

A Figura 2 ilustra as etapas do processo de classificação de mensagens eletrônicas. O vetor de características é utilizado como entrada para o sistema computacional responsável pelo processo de classificação. Para a função de agente classificador são empregadas diferentes técnicas de aprendizado de máquina e, neste trabalho, são utilizadas as redes neurais artificiais. Em seguida, é feita uma análise dos resultados da classificação com o objetivo de verificar seu desempenho. Nas próximas seções as etapas descritas anteriormente serão detalhadas.

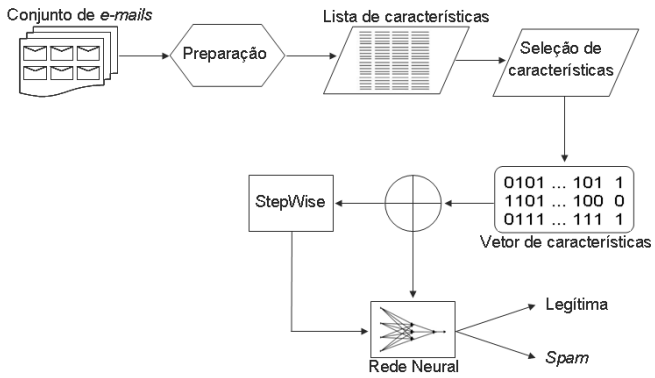

Figure 2: Processo de classificação de mensagens eletrônicas.

4.1 Conjunto de Dados

A definição de um bom conjunto de dados, com mensagens representativas das categorias definidas, é de grande importância para o sucesso na classificação. Devido à existência de diversos tipos de mensagens legítimas e *spams*, é importante definir um conjunto de dados que contemple satisfatoriamente todos os tipos.

Na internet estão disponíveis diversos *Corpora* de mensagens que são utilizados pela comunidade científica em seus experimentos, dentre eles: *Ling-Spam* (Androutsopoulos et al., 2000); PU1 e PU123 (Clark et al., 2003); *Eron-Spam* (Mildinhall and Noyes, 2008); *SpamAssassin* (SpamAssassin, 2008).

Neste trabalho, optou-se por utilizar o conjuntos de mensagens do *SpamAssassin*. Diferente de outros conjuntos de mensagens como PU1, PU123A e *Ling-Spam* no do *SpamAssassin* as mensagens estão em seu formato original, sem remoção de *tags* HTML, anexos ou conteúdos dos *e-mails*, o que é essencial para a realização deste trabalho, já que este se propõe a utilizar todo o conteúdo da mensagem.

A grande utilização da base do *SpamAssassin* em trabalhos relacionados é importante para servir como base de comparação para os resultados alcançados e apresentam um bom número de *e-mails*, conseguindo representar bem as categorias de mensagens legítimas e *spam*, sendo de grande importância para um treinamento representativo, validação e teste da rede neural. Este conjunto de mensagens é descrito na próxima seção.

4.1.1 *SpamAssassin Public Corpus*

O *Corpus* público de *e-mails* do *SpamAssassin* é uma coleção de mensagens criada especialmente para o uso em testes de sistemas anti-*spam*. Esse conjunto de mensagens é amplamente utilizado pela comunidade científica em seus experimentos, como nas pesquisas de Bergholz et al. (2008a), Bergholz et al. (2008b), Sculley and Cormack (2008), Mojdeh and Cormack (2008), Assis (2006), Carpinteiro et al. (2006) e Rios and Zha (2004). É composto por três[5] conjuntos de mensagens, como segue:

1. ***Spam***: 1.897 mensagens de *spams*, recebidas de não fontes de *spam*. Inicialmente, esse conjunto era composto por 500 mensagens, posteriormente foram adicionadas outras 1.397 ao conjunto;

2. ***Easy Ham***: 3.900 mensagens legítimas, facilmente diferenciáveis de *spams* por não conterem nenhuma assinatura de *spam* como HTML, por exemplo. Esse conjunto era composto por 2.500 mensagens, as outras 1.400 foram incorporadas a *posteriori*;

3. ***Hard Ham***: 250 mensagens legítimas, porém mais difíceis de diferenciar dos *spams*. Essas fazem uso de HTML, texto colorido etc.

O *Corpus* possui um total de 6.047 mensagens, sendo aproximadamente 31% de *spams*.

4.2 Preparação das Mensagens

O objetivo desta etapa é realizar uma uniformização do conteúdo das mensagens, transformando as informações complexas presentes em cada mensagem, em informações mais simples, permitindo um melhor desempenho na classificação. Nesse processo, todos os *e-mails* do conjunto são analisados e uniformizados. Várias técnicas de preparação de textos podem ser empregadas, entre elas a remoção de *stopwords* e a técnica de *stemming*.

Nem todas as palavras são igualmente significativas para representar uma categoria, algumas, por exemplo, carregam mais significado que outras. Normalmente os substantivos, seguidos dos adjetivos e verbos carregam mais representatividade do que outras classes gramaticais como os pronomes, conjunções e artigos. A remoção das *stopwords* consiste em descartar as palavras que pouco refletem o conteúdo de um documento ou são tão comuns que não distinguem nenhuma categoria dos documentos.

Uma determinada palavra em um texto pode assumir formas variadas, como por exemplo: inteligência, inteligente, inteligentemente. O *Stemming* consiste em remover os prefixos e sufixos. O resultado obtido é chamado de *Stem* (raiz ou radical) (Monteiro et al., 2006).

Nesta pesquisa, emprega-se o método de preparação proposto por Assis (2006) e Carpinteiro et al. (2006), no qual todos os caracteres são convertidos em minúsculos; imagens, anexos, *links*, endereços eletrônicos, moeda, porcentagem e palavras longas são substituídos por *strings* específicos; os acentos são removidos; *tags* HTML são tratadas, sendo que algumas são utilizadas integralmente e outras parcialmente ou descartadas; palavras pequenas são descartadas; entre outras.

Para entender o processo de uniformização se faz necessário apresentar a estrutura do *e-mail*, tal como é feito a seguir. Já na seção subsequente da apresentação da estrutura do *e-mail* será descrito o processo de uniformização.

[5]Foram utilizados os *Corpus* de mensagens com as numerações 20021010, 20030228 e 20050311.

4.2.1 Estrutura do *E-mail*

A mensagem de correio eletrônico é definida pela RFC 822[6] (*Internet Message Format*) e é composta basicamente de duas partes: cabeçalho (*header*) e corpo (*body*). O cabeçalho contém as informações do protocolo utilizado, do remetente, data, hora, assunto, domínios, entre outras. Cada uma dessas informações é um campo.

O corpo do *e-mail* é a mensagem propriamente dita, podendo conter texto plano, HTML ou anexos. Uma mensagem também pode conter vários tipos de conteúdo simultaneamente, sendo que o padrão que possibilita essa propriedade é o MIME (*Multipurpose Internet Mail Extensions*), descrito nas RFC 2045[7], 2046[8], 2047[9], 2048[10] e 2387[11]. O cabeçalho MIME (*MIME-Version: 1.0*) é incluído antes de cada conteúdo MIME e o campo *Content-Type* é responsável por identificar o tipo (*type*) e o subtipo (*subtype*) do conteúdo presente no *e-mail*. A seguir são apresentados exemplos desse campo.

1. *Content-type: text/plain*
2. *Content-type: image/jpeg*
3. *Content-type: application/octet-stream*
4. *Content-type: multipart/alternative*

4.3 Uniformização

A proposta de uniformização de Carpinteiro et al. (2006) e Assis (2006) visa aproveitar quase toda informação presente nos *e-mails*. Esse processo pode ser separado em três diferentes etapas: processamento HTML, tokenização e detecção de padrões.

A Figura 3 ilustra o processo de uniformização, onde é verificado, através do campo *Content-Type* do cabeçalho, o tipo de informação presente na mensagem. Caso seja texto plano, é enviado diretamente para a tokenização, e, se HTML, é enviado para o processamento HTML. O campo *Content-Type* também pode ser *multipart-alternative*, indicando que a mensagem possui informações em texto plano e em HTML. Nessa situação, o texto plano é descartado e somente o conteúdo HTML é processado.

Esse procedimento se deve ao fato de que somente a mensagem HTML é visualizada no leitor de *e-mails* e os *spammers*, conforme já dito, utilizam a técnica para enviar no texto plano uma mensagem legítima e o texto de *spam* no conteúdo HTML. Ao final desse processamento o conteúdo HTML é transformado em texto plano e enviado para a tokenização. Após a tokenização a mensagem é enviada para a detecção de padrões. A seguir os procedimentos supracitados serão descritos em maiores detalhes.

4.3.1 Processamento HTML

HTML é uma linguagem interpretada que possibilita adicionar ao *e-mail* formatação de texto, inclusão de tabelas, *hyperlinks* e imagens como em qualquer página da *web*. Essas informações

[6] http://www.ietf.org/rfc/rfc2822.txt
[7] http://www.ietf.org/rfc/rfc2045.txt
[8] http://www.ietf.org/rfc/rfc2046.txt
[9] http://www.ietf.org/rfc/rfc2047.txt
[10] http://www.ietf.org/rfc/rfc2048.txt
[11] http://www.ietf.org/rfc/rfc2387.txt

são adicionadas por meio de *tags*, que são rótulos usados para informar ao aplicativo como as informações devem ser apresentadas. Todas as *tags* têm o mesmo formato: <nome-*tag* parâmetros> Texto da *tag* <nome-*tag*>. O que estiver contido entre uma *tag* de abertura e uma *tag* de fechamento será processado segundo o comando contido na *tag*.

Para o melhor aproveitamento da informação presente no conteúdo, as *tags* foram divididas em três categorias de acordo com o grau de importância para o processo, sendo que cada uma destas possui um tipo de processamento especial. Tal como se segue:

1. Nessa categoria, toda informação da *tag* é descartada, incluindo seus parâmetros e comentários. Assis (2006) supôs que as informações presentes nessas *tags* são irrelevantes.

2. As *tags* desta categoria são substituídas por outra específica, composta pelos caracteres *!_in_* mais a *tag*, os seus atributos são descartados. A *tag* <*label for=e-mail*>*e-mail address*</*label*> é substituída por *!_in_label e-mail address*, por exemplo.

3. Estas *tags* são processadas integralmente. A *tag*, seus parâmetros e conteúdo são adicionados à saída. A *tag* é substituída por outra específica, como na categoria 2. O conteúdo da *tag* também é pré-processado e adicionado à saída, como no exemplo a seguir: o código HTML <*form action=result.php*> *Conteúdo* </*form*> é transformado em *!_inform action conteúdo* durante o processamento.

4.3.2 Tokenização

A tokenização é o processo de decompor a mensagem em cada termo que a compõe, sendo esse termo conhecido como *token*. A tokenização separa o texto em palavras soltas. Este processo é executado nas mensagens com texto plano e nas mensagens com conteúdo HTML após o processamento. Os delimitadores utilizados por Assis (2006) para a tokenização foram: espaço; nova linha; tabulação; exclamação; interrogação; vírgula; ponto-e-vírgula.

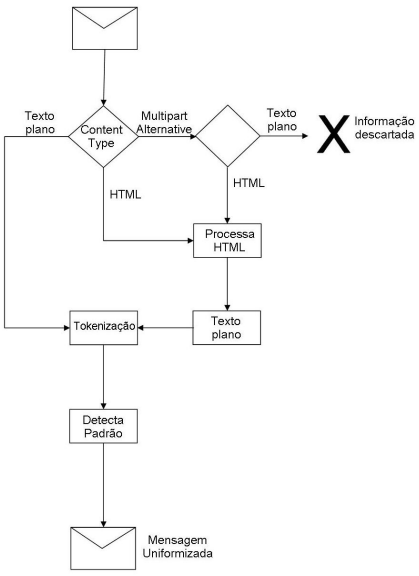

Figure 3: Esquema de funcionamento do processo de uniformização.

Para melhorar a uniformização das informações presentes em cada mensagem, após a tokenização é realizada a transformação de todos os caracteres em minúsculos e a acentuação é removida. Após esse processo, os *tokens* são enviados para a detecção de padrões.

4.3.3 Detecção de Padrões

A detecção de padrões é utilizada para ajudar na identificação de padrões típicos empregados pelos *spammers* e para unificar outros padrões de saída. Os seguintes padrões são detectados:

- valores de saída de parâmetro HTML são unificados. Exemplo: o campo *table color=red* é transformado em *!_table color*;
- as referências a *e-mail*, URL ou *hyperlinks* são uniformizadas. Exemplo: o endereço eletrônico *teste@subdominio.cefetmg.br* é substituído por *!_e-mail*, o mesmo acontece com *hyperlinks* que são convertidos em *!_link*;
- os caracteres encontrados no meio das palavras são detectados e qualquer palavra com esse padrão é substituída por *!_HIDEWORD*;
- palavras muito grandes, acima de 20 caracteres são substituídas por *!_BIGTEXT*. Geralmente são palavras sem sentido e utilizadas com o objetivo de enganar os filtros;
- o uso de números ou *strings* inválidas no sujeito (*subject*) é uma das técnicas utilizadas pelos *spammers* para enganar os filtros anti-*spam*. Essas *strings* são detectadas e substituídas pela *string !_NUMERO_SUBJECT*. Exemplo: a mensagem *Get the best Life can ofter you @e90jaakdfd* no sujeito do *e-mail* é transformada em *get the best life can ofter you !_NUMERO_SUBJECT*;
- as ocorrências de unidades monetária e porcentagem são transformadas nas *string !_MONEY* e *!_PORCENTAGEM*, respectivamente;
- os anexos são descartados e substituídos por *!_ANEXO_TIPO?nome_anexo*. Exemplo: uma figura formato jpeg é substituída por *!_ANEXO_TIPO?image/jpeg*.

Assim, com a detecção dos padrões e uniformização das informações contidas nas mensagens, estas estão prontas para serem utilizadas para gerar o conjunto de estatísticas. Esse processo é conhecido como seleção de características. Para cada palavra é calculado um valor em cada um dos métodos estatísticos.

4.4 Seleção de Características

O principal problema enfrentado pelas diferentes técnicas de categorização de texto é a alta dimensionalidade do espaço característico (Yang and Pedersen, 1997). Em um conjunto de dados uma característica é uma palavra e o espaço característico é no número total de palavras contidas nos documentos, que podem ser dezenas ou milhares, variando de acordo com a quantidade de documentos e informações contidas nestes. No conjunto de mensagens do *SpamAssassin* o espaço característico é de 56.646 palavras.

A alta dimensionalidade do espaço característico torna a categorização de texto inviável devido ao alto custo computacional, independentemente da técnica empregada. Por isso, é altamente desejável reduzir o espaço característico. É recomendável que essa redução seja

realizada de forma automática e, principalmente, sem sacrificar a precisão na classificação dos documentos. A técnica empregada nesse processo é a da seleção de características.

Segundo Junior (2004) a redução de dimensionalidade busca identificar um subespaço suficientemente reduzido de características, que seja capaz de representar qualquer padrão conhecido de acordo com um determinado critério. É recomendável que essa redução seja realizada de forma automática e, principalmente, sem sacrificar a precisão na classificação.

Basicamente, duas abordagens podem ser empregadas para tratar esse problema: a fusão de características e a seleção de características (Junior, 2004). Os algoritmos de fusão de características criam novas características a partir de transformações ou combinações do conjunto original, um exemplo deste tipo é o PCA (*Principal Analysis Components* - Análise dos Componentes Principais). A seleção de características consiste na utilização de métodos estatísticos na extração das informações mais relevantes de um conjunto de dados, identificando os dados que melhor representam uma categoria. Estes algoritmos analisam cada característica individualmente e seleciona as n melhores. Nesta pesquisa serão empregados os métodos de seleção de características Distribuição por Frequência, Informação Mútua e χ^2 *statistic*, que serão detalhados nas próximas seções.

4.4.1 Distribuição por Frequência (DF)

A DF é uma das técnicas mais simples para redução da dimensionalidade. Possui uma complexidade computacional aproximadamente linear, o que possibilita seu uso em grandes conjuntos de dados a um custo computacional relativamente pequeno.

A escolha desse método se deu pelo seu desempenho nos experimentos de Carpinteiro et al. (2006); Assis (2006); Drucker et al. (1999); Yang and Pedersen (1997) e também por ser o mais simples dos métodos necessitando assim, menores recursos computacionais. Outro motivador é a possibilidade de utilizar variações nos cálculos que não foram implementadas nos trabalhos supracitados.

A distribuição por frequência é definida pelo número de ocorrência de um termo em um conjunto de elementos (Yang and Pedersen, 1997). O cálculo da DF de uma palavra se dá por meio da Equação 1:

$$DF = \frac{N[x \in legitima, spam]}{T}, \tag{1}$$

onde N é o número de ocorrência da palavra x na classe $legitima, spam$ e T o número total de palavras na classe.

Para representar as classes são escolhidas as palavras com valores de DF mais altos, considerando-se que as palavras com baixa frequência de ocorrência são menos significativas para a identificação das classes.

Nessa técnica, a palavra possui um valor de DF para cada uma das classes, de modo que um DF para o conjunto de mensagens legítimas e outro para o de *spams*. A proposta deste procedimento é variar a forma de cálculo do DF unificado da palavra (DF Legítimo e DF *Spam*). O objetivo dessas diferentes formas de cálculo é conseguir encontrar as características que melhor definam cada uma das classes e, consequentemente, melhorem o desempenho na classificação. A seguir são apresentadas as três variações implementadas nesta pesquisa.

DF SOMA ($DF+$) O cálculo comumente utilizado pela comunidade científica para encontrar o DF de uma palavra é somar o seu DF em cada uma das classes, como pode ser visto

na Equação 2. As palavras com valores de DF mais altos são selecionadas para compor o vetor de características. Esse método foi empregado no trabalho de Carpinteiro et al. (2006) e Assis (2006).

$$DF+ = DF\_Legitimo + DF\_Spam \tag{2}$$

DF Exclusão de Termos Comuns ($DFETC$) Algumas palavras possuem um elevado número de ocorrência nas duas classes e, assim, um alto valor de DF em ambas. Baseado no princípio de que se uma palavra possui alta representatividade em duas classes distintas não será uma boa representante para uma classe específica.

Diante dessa premissa, este método busca excluir as palavras com elevado valor de DF que estão presentes nas duas classes, selecionando apenas as palavras que possuem elevado DF em uma única classe. Por exemplo, num conjunto de 100 palavras de cada uma das classes, são selecionadas para compor o vetor de características, aquelas que estão presentes somente em uma das classes.

DF Subtração ($DF-$) Nesse método são selecionadas as palavras que possuem a maior diferença entre o seu valor de DF nas duas classes. O DF é obtido através do módulo (valor absoluto) da subtração entre os valores de DF da palavra nas duas classes. Após o cálculo descrito na Equação 3, as palavras com maior DF são selecionadas para compor o vetor de características.

$$DF- = |DF\_Legitimo - DF\_Spam| \tag{3}$$

4.4.2 Informação Mútua - *Mutual Information* (MI)

Informação Mútua é um método estatístico amplamente utilizado em categorização de texto para redução de dimensionalidade (Chuan et al., 2005). Sua escolha para emprego neste trabalho se deu pela sua ampla utilização no processo de redução de dimensionalidade, bem como por seu bom desempenho nos trabalhos de Carpinteiro et al. (2006); Assis (2006); Chuan et al. (2005); Ozgur et al. (2004); Androutsopoulos et al. (2000).

Sendo w uma característica, o MI de w é dado pela Equação 4:

$$MI(w) = \sum_{w\in\{0,1\},c\in\{legitimo,spam\}} P(W=w, C=c).\log\frac{P(W=w, C=c)}{P(W=w).P(C=c)} \tag{4}$$

onde $c = classe(legitimo, spam)$, $P(W=w, C=c)$ é a probabilidade que a palavra w ocorra $(w=1)$ ou não ocorra $(w=0)$ em *spam* $(c=spam)$ ou legítimo $(c=legitimo)$, $P(W=w)$ é a probabilidade que a palavra w ocorra em todos os *e-mails* e $P(C=c)$ é a probabilidade de um *e-mail* ser *spam*. As palavras com valores mais altos de MI são selecionadas.

4.4.3 χ^2 *statistic* (QUI)

O QUI mede a independência entre t e C, onde t é um elemento e C um conjunto (Yang and Pedersen, 1997). A distribuição QUI para uma característica w e uma classe c é dada pela Equação 5:

$$QUI(w,c) = \frac{N.(Kn-ml)^2}{(k+m).(l+n).(K+l)(m+n)} \tag{5}$$

onde k é o número de *e-mails*, dentro da classe c, que contém a característica w. l é o número de *e-mails*, dentro da classe $\bar{c}$, que contém a característica w. m é o número de *e-mails*, dentro da classe c que não contém a característica w. n é o número de *e-mails*, dentro da classe $\bar{c}$ que não contém a característica w, e N é o número total de *e-mails* dentro da classe c.

A distribuição QUI de uma característica t dentro de um conjunto C com duas classes $(legitimo, spam)$ é dada pela Equação 6:

$$QUI(t) = P(spam).QUI(t, spam) + P(legitimo).QUI(t, legitimo) \quad (6)$$

onde $P(spam)$ e $P(legitimo)$ são as probabilidades da ocorrência de *e-mails* spam e legítimos respectivamente. As características com os valores mais altos de QUI são escolhidas. Cada característica é uma entrada para o agente classificador. A escolha desse método se deu pelos resultados apresentados nos experimento de Assis (2006), Meyer and Whateley (2004) e Yang and Pedersen (1997).

4.5 Vetor Característico

Empregando os métodos de seleção de características apresentados acima, é calculado para cada palavra o seu valor. Um arquivo com as estatísticas de cada palavra para cada um dos métodos é gerado, sendo selecionadas as características que irão compor o vetor de características.

O vetor característico é criado a partir da seleção das n características mais relevantes de acordo com o método empregado. Para os experimentos foram gerados vetores com 25 e 50 características para os métodos de extração supracitados. Cada característica corresponde a um nó de entrada da RNA e cada mensagem é representada por um vetor $X = (x_1, x_2, ..., x_n)$, onde n é o número de características. Em testes preliminares foram analisadas as seguintes técnicas para compor o vetor: Frequência do Termo; Peso Binário; Peso Normal. O método do peso binário apresentou, nos testes realizados, os melhores resultados na classificação e o menor consumo de tempo e recursos computacionais durante o processamento. O peso binário foi empregado por Assis (2006) e Carpinteiro et al. (2006). A Figura 4 ilustra um vetor de 19 características, nesta figura cada linha corresponde a uma mensagem e as colunas indicam as características, sendo que última identifica a classe do *e-mail* (0 para mensagem legítima e 1 para *spam*).

5 *STEPWISE REGRESSION*

As técnicas de seleção de variáveis buscam determinar qual o melhor subconjunto de variáveis de entrada para compor um modelo. A técnica *StepWise* (passo a passo) utiliza uma técnica de regressão linear múltipla para escolha de variáveis. O modelo começa com todas as variáveis do conjunto e remove de forma gradativa as que são estatisticamente menos significantes. Esse processo ocorre até que as variáveis restantes sejam todas importantes (estatisticamente relevantes), ou seja, até que não haja melhora no desempenho do modelo ou não haja variáveis a serem retiradas. Essa técnica supõe que algumas variáveis não contribuem de forma significativa para a resposta de todo o conjunto (Demuth et al., 2008). Após a retirada de uma variável, esta não poderá mais compor o modelo.

Segundo Junior (2004) a aplicação dessas técnicas pode facilitar o trabalho de modelagem e melhorar os resultados obtidos. Em estudo comparativo entre vários métodos para seleção de variáveis em modelagem RNA do vapor gerado por uma caldeira realizado por Meireles et al. (2003), o método de *StepWise* foi um dos que apresentou os melhores resultados.

Nestes experimentos a técnica *StepWise* foi empregada nos vetores de entrada das redes neurais com o objetivo de reduzir a sua dimensionalidade. A Tabela 1 ilustra o resultado dessa redução para o conjunto do *SpamAssassin*. Nelas a primeira coluna identifica o método de seleção de características, da segunda a quarta colunas são apresentados os resultados da redução de dimensionalidade obtida após o uso de *StepWise* para 25 e 50 elementos no vetor de entrada da rede.

Table 1: Resultado do uso de *Stepwise* na base do *SpamAssassin*

Método Sel. Características	StepWise 25	StepWise 50
DF ETC	19	36
DF+	14	31
DF-	20	38
QUI	23	41
MI	22	36

6 METODOLOGIA

Nos testes realizados, foram empregados os método de seleção de características: Distribuição por Freqüência (três variações); Informação Mútua e; χ^2 *statistic*. Como agente classificador foram utilizadas as redes neurais MLP (*MultiLayer Perceptron*) com uma única camada intermediária (5, 25 ou 50 neurônios). O algoritmo de treinamento utilizado foi o *Levenberg-Marquardt*. A técnica de validação cruzada, que objetiva encerrar o treinamento quando o erro de validação começa a aumentar, indicando que a rede está aprendendo com o ruído presente nos dados, foi implementada neste trabalho.

Segundo Haykin (2001), na validação cruzada o conjunto de dados é dividido aleatoriamente em um conjunto de treinamento, teste e validação. O conjunto de treinamento é utilizado no

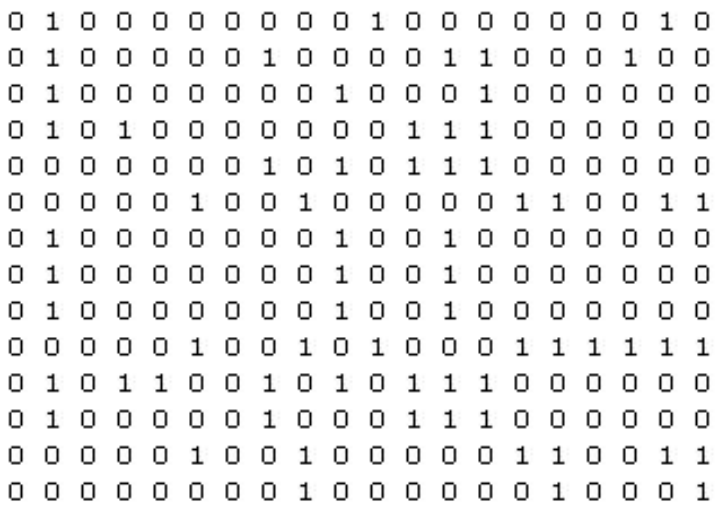

Figure 4: Vetor de características indexado por peso binário.

processo de aprendizagem da rede. O de teste serve para interromper a aprendizagem, quando o erro médio quadrático para os dados de teste começa a aumentar de forma contínua. Os dados de validação são utilizados para verificar a capacidade de generalização da rede. O conjunto de dados foi distribuído da seguinte forma: 60% para treinamento, 20% para teste e 20% para validação (Demuth et al., 2008).

Cada experimento foi executado 30 vezes e na avaliação dos resultados de cada configuração de rede foram considerados a média e o melhor resultado das execuções do experimento.

6.1 Medidas de Desempenho

O desempenho na classificação e categorização pode ser avaliado através de diferentes métodos. Esses podem ir desde o cálculo da abrangência e precisão até o cálculo da acurácia (*accuracy*), sensitividade (*recall*), *ranking*, micro e macro-média, entre outros (Yang and Pedersen, 1997). De acordo com Androutsopoulos et al. (2000), em tarefas de classificação, o desempenho é frequentemente mensurado pelo processo de acurácia ou pela taxa de erro (Err).

As letras L e S foram usadas para *e-mails* legítimos e *spams*, respectivamente, n_L e n_S como o número total de mensagens legítimas e *spams*, $n_{L\to L}$ e $n_{S\to S}$ como o total de mensagens legítimas e *spams* classificados corretamente. O número de mensagens legítimas classificadas incorretamente como *spams* é indicado por $n_{L\to S}$ e o de *spams* classificados como mensagens legítimas por $n_{S\to L}$. Neste trabalho será utilizada o Erro que indica a proporção de classificações incorretas, e é dado pela Equação 7:

$$Erro = \frac{n_{S\to L} + n_{L\to S}}{n_S + n_L} \tag{7}$$

7 RESULTADOS OBTIDOS

Os resultados dos experimentos com 25 neurônios na camada de entrada (com e sem o uso da *StepWise*) são ilustrados nas Figuras 5 e 6. A Figura 5 apresenta o percentual do erro médio e a Figura 6 o percentual dos menores erros obtidos em cada método de seleção de características. Pode-se perceber que com 25 elementos de entrada para o classificador o uso da técnica *StepWise* conseguiu, além de reduzir o número de características, reduzir a taxa de erro.

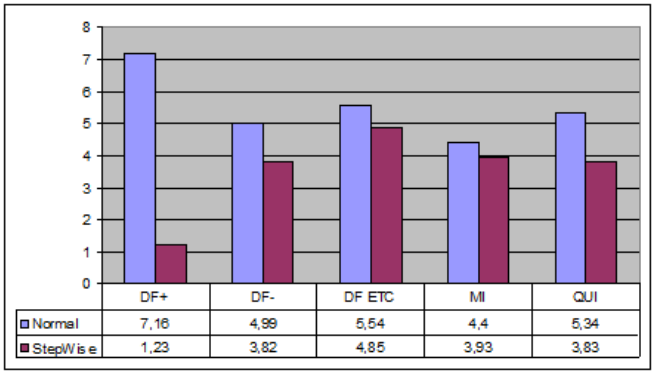

Figure 5: Percentual de erro médio com 25 elementos de entrada.

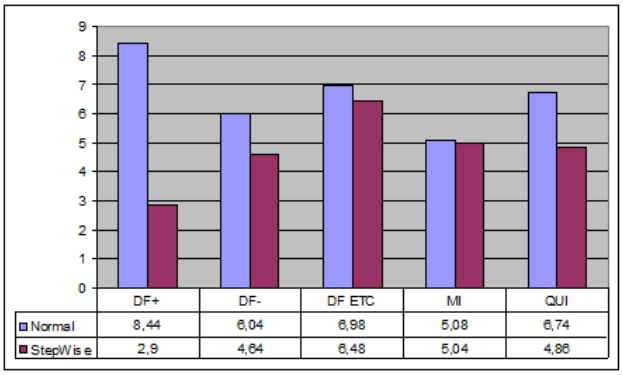

Figure 6: Menor percentual de erro com 25 elementos de entrada.

As Figuras 7 e 8 ilustram os resultados obtidos nos experimentos com 50 características. As médias dos erros são apresentados na Figura 7 e as menores taxas de erros de cada experimento na Figura 8. Nestes experimentos os resultados do uso da técnica *StepWise* só não foram superiores quando empregada em conjunto com o método de seleção de características DF Soma, isto aconteceu tanto na média quanto na menor taxa de erro.

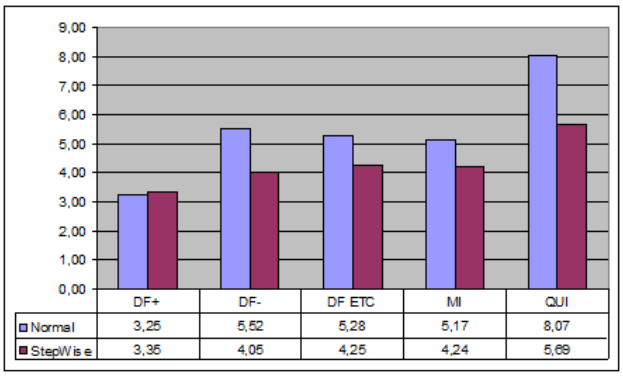

Figure 7: Percentual de erro médio com 50 elementos de entrada.

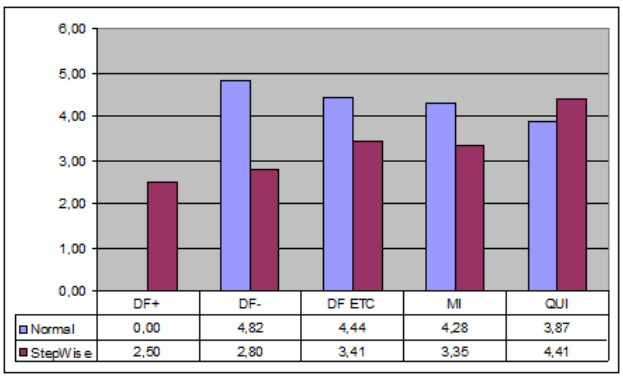

Figure 8: Menor percentual de erro com 50 elementos de entrada.

8 CONSIDERAÇÕES FINAIS

Este trabalho apresentou um sistema para classificação de *e-mails* utilizando redes neurais artificiais. A técnica de regressão linear múltipla *StepWise* foi empregada em conjunto com métodos de seleção de características para reduzir a dimensionalidade do vetor de entrada para o classificador. Com os experimentos realizados e os resultados obtidos pode-se verificar que a *StepWise* além de reduzir o número de características conseguiu melhorar o desempenho do classificador. Somente em um dos casos (50 elementos de entrada com o método de seleção de características DF Soma) o desempenho da técnica *StepWise* não foi superior.

AGRADECIMENTOS

Agradecemos a FAPEMIG pelo apoio financeiro concedido.

REFERENCES

Andrade L.M. *Análise Comparativa de Técnicas de Inteligência Computacional para a Detecção de Spam*. Master's Thesis, UFMG - Programa de Pós-Graduação em Engenharia Elétrica, 2006.

Androutsopoulos I., Koutsias J., Konstantinos C., and Spyropoulos C. An experimental comparison of naive bayesian and keyword-based anti-spam filtering with personal e-mail messages. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 160–167. ACM, New York, NY, USA, 2000. ISBN 1-58113-226-3. doi:http://doi.acm.org/10.1145/345508.345569.

Assis J.M.C. *Detecção de E-mails Spam Utilizando Redes Neurais Artificiais*. Master's Thesis, Universidade Federal de Itajubá - Programa de Pós-Graduação em Engenharia Elétrica, 2006.

Bergholz A., Chang J.H., Paass G., Reichartz F., and Strobel S. Improved phishing detection using model-based features. In *Proceedings of the Fifth Conference on Email and Anti-Spam*. CEAS, Mountain View, CA, USA, 2008a.

Bergholz A., Paass G., Reichartz F., Strobel S., Moens M.F., and Witten B. Detecting known and new salting tricks in unwanted emails. In *Proceedings of the Fifth Conference on Email and Anti-Spam*. CEAS, Mountain View, CA, USA, 2008b.

Carpinteiro O.A.S., Lima I., J. M. C. Assis A.C.Z.S., Moreira E.M., and Pinheiro C.A.M. A neural model in anti-spam systems. In *Artificial Neural Networks - ICANN 2006, 16th International Conference*, volume 4132 of *Lecture Notes in Computer Science*, pages 847–855. Springer, Athens, Greece, 2006. ISBN 3-540-38871-0.

Carreras X. and Marquez L. Boosting trees for clause splitting. In *ConLL '01: Proceedings of the 2001 workshop on Computational Natural Language Learning*, pages 73 – 75. Association for Computational Linguistics, Morristown, NJ, USA, 2001.

Chuan Z., Xianliang L., Mengshu H., and Xu Z. A lvq-based neural network anti-spam email approach. *SIGOPS Operating Systems Review*, 39(1):34 – 39, 2005. ISSN 0163-5980. doi:http://doi.acm.org/10.1145/1044552.1044555.

Clark J., Koprinska I., and Poon J. Linger - a smart personal assistant for e-mail classif. *ICANN/ICONIP*, 2003.

Cormack G. and Lynam T. Spam corpus creation for trec. In *Proceedings of the Second Conference on Email and Anti-Spam*. CEAS, Mountain View, CA, USA, 2005.

Cranor L.F. and LaMacchia B.A. Spam! *Commun. ACM*, 41:74–83, 1998. ISSN 0001-0782.

Demuth H., Beale M., and Hagan M. *Neural Network Toolbox 6*. The MathWorks, Natic, MA, USA, 2008.

Drucker H., Wu D., and Vapnik V.N. Support vector machines for spam categorization. *IEEE Transactions on Neural Networks*, 10(5):1048 – 1054, 1999.

Evett D. Spam statistics. Site: Spam Filter, 2006.

Guzella T.S., Uchôa J.Q., Santos T.A.M., and Caminhas W.M. Proposta de um modelo de classificação de padrões baseado no sistema imune: uma aplicação para a identificação de spam. In *VII Congresso Brasileiro de Redes Neurais*, volume 1. Anais do VII Congresso Brasileiro de Redes Neurais, Natal, RN, Brasil, 2005.

Haykin S. *Redes Neurais*. Bookman, Porto Alegre, 2ł edition, 2001.

IronPort. Ironport study on email authentication reveals significant adoption. Site: Home page, 2006.

IronPort. Trends report ironport. Site: Home page, 2008.

Junior D.C.M. *Redução de dimensionalidade utilizando entropia condicional média aplicada a problemas de bioinformática e de processamento de imagens*. Master's Thesis, USP, 2004.

Kaspersky L. Introduction to spam. Site: Home page, 2008.

Kosmopoulos A., Paliouras G., and Androutsopoulos I. Adaptive spam filtering using only naive bayes text classifiers. In *Proceedings of the Fifth Conference on Email and Anti-Spam*. CEAS, Mountain View, CA, USA, 2008.

Levine J.R. Experiences with greylisting. In *Proceedings of the Second Conference on Email and Anti-Spam*. CEAS, Mountain View, CA, USA, 2005.

Marshal. Spam volumes hit record high. Site: New Room, 2007.

Marshal. Threat research and content engineering. Site: Trace, 2008.

Meireles M.R.G., Almeida P.E.M., and Simões M.G. A comprehensive review for industrial applicability of artificial neural networks. *IEEE Transactions on Industrial Electronic*, 50(3):585–601, 2003.

Meyer T.A. and Whateley B. Spambayes: Effective open-source, bayesian based, e-mail classification systems. In *Proceedings of the First Conference on Email and Anti-Spam*. CEAS, Mountain View, CA, USA, 2004.

Mildinhall J. and Noyes J. Toward a stochastic speech act model of email behavior. In *Proceedings of the Fifth Conference on Email and Anti-Spam*. CEAS, Mountain View, CA, USA, 2008.

Mojdeh M. and Cormack G. A mail client plugin for privacy-preserving spam filter evaluation. In *Proceedings of the Fifth Conference on Email and Anti-Spam*. CEAS, Mountain View, CA, USA, 2008.

Monteiro L.O., Gomes I.R., and Oliveira T. Etapas do processo de mineração de textos: uma abordagem aplicada a textos em português do brasil. In *WCOMPA I Workshop de Computação e Aplicações*. Anais do XXVI Congresso da Sociedade Brasileira de Computação, Campos Grande, MS, Brasil, 2006.

Nucleus R.I. Spam: The silent roi killer. *Research Note NOTE D59*, 2003.

Nucleus R.I. Spam: The repeat offender. *Research Note Document H22*, 2007.

Ozgur L., Gungor T., and Gurgen F. Adaptive anti-spam filtering for agglutinative languages: a special case for turkish. *Pattern Recognition Letters*, 25(16):1819 – 1831, 2004. ISSN 0167-8655. doi:http://dx.doi.org/10.1016/j.patrec.2004.07.004.

Rios G. and Zha H. Exploring support vector machines and randon forest for spam detection. In *Proceedings of the First Conference on Email and Anti-Spam*. CEAS, Mountain View, CA, USA, 2004.

Rizzi C.B., Wives L.K., de Oliveira J.P.M., and Engel P.M. Fazendo uso da categorização de textos em atividades empresariais. In *Proceedings of the International Symposium on*

Knowledge Management/Document Management. PUC-PR, Curitiba, PR, Brasil, 2000.

Sakkis G., Androutsopoulos I., Paliouras G., Karkaletsis V., Spyropoulos C.D., and Stamatopoulos P. A memory-based approach to anti-spam filtering for mailing lists. *Information Retrieval*, 6(1):49 – 73, 2003. ISSN 1386-4564.

Sculley D. and Cormack G. Filtering e-mail spam in the presence of noisy user feedback. In *Proceedings of the Fifth Conference on Email and Anti-Spam*. CEAS, Mountain View, CA, USA, 2008.

SpamAssassin. The apache spamassassin project - the apache software foundation. Site: Home page, 2008.

Symantec. The state of spam: A monthly report - august 2008. Site: Home page, 2008.

Teixeira R.C. *Combatendo o Spam*. Novatec, São Paulo, 1ł edition, 2004.

Trudgian D.C. and Yang Z. Spam classification using nearest neighbour techniques. In *Proc. of Fifth International Conf. on Intelligent Data Engineering and Automated Learning*, volume 3177, pages 578 – 585. Springer, Berlin, Alemanha, 2004. ISBN 978-3-540-22881-3.

Wittel G.L. and Wu S.F. On attacking statistical spam filters. In *Proceedings of the First Conference on Email and Anti-Spam*. CEAS, Mountain View, CA, USA, 2004.

Yang Y. and Pedersen J.O. A comparative study on feature selection in text categorization. In *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*, pages 412–420. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997. ISBN 1-55860-486-3.

Zhang L., Zhu J., and Yao T. An evaluation of statistical spam filtering techniques. *ACM Transactions on Asian Language Information Processing (TALIP)*, 3(4):243–269, 2004. ISSN 1530-0226. doi:http://doi.acm.org/10.1145/1039621.1039625.