Classificação Supervisionada de Mensagens Eletrônicas Indesejadas Utilizando os Algoritmos MLP e FBR

Khalil Carsten do Nascimento
Departamento de Ciência da Computação
Universidade de Brasília
Brasília, Brasil
khalilcarsten@gmail.com

Renato Avelar Nobre
Departamento de Ciência da Computação
Universidade de Brasília
Brasília, Brasil
rekanobre@gmail.com

Index Terms*—rede de função de base radial, perceptron multicamadas, *e-mail*, *spam

I. Introdução

O número de mensagens trocadas pelo correio eletrônico, *e-mails*, diariamente excedem 144,8 bilhões enviados mundialmente. No entanto, 65% deste valor consiste de mensagens não solicitadas ou esperadas pelo usuário, consistindo em sua maioria de assuntos publicitários ou tráfico de informações falsas. Estes tipos de *e-mails* são denominados de *spam*.

A alta quantidade de *spam*, além de encher a caixa de mensagens do usuário com conteúdo indesejado, desviando o foco do mesmo dos *e-mails* que realmente são importantes, pode conter conteúdo com o objetivo de cometer estelionato. Logo, percebe-se uma necessidade de realizar um filtro de *e-mails* antes mesmo que chegue ao destinatário.

No entanto, a detecção deste tipo de conteúdo não é um processo trivial, visto que a detecção do mesmo necessita de um processamento de linguagem natural para interpretar diversas áreas da mensagem onde podem haver indicativos de conteúdo indesejado. Portanto, este trabalho propõe uma abordagem utilizando algoritmos de aprendizagem de máquina para tentar classificar suspeitas de *spam*. As abordagens foram realizadas com os algoritmos de Perceptron Multicamadas, e Rede de Função de Base Radial.

II. Especificações do Sistema

A estrutura do sistema foi desenvolvida para evitar acoplamento entre os algoritmos de aprendizagem, os algoritmos de tratamento de imagens, e os algoritmos de suporte. O sistema adaptador é responsável por conciliar o ambiente para respeitar as interfaces requeridas pelo usuário ao realizar entradas pela linha de comando.

Todo o sistema foi construído utilizando a linguagem de programação *Python 3* que é a linguagem usada em milhares de aplicações de negócio no mundo, incluindo diversos sistemas de larga escala [1]. Para o auxilio no desenvolvimento das funções de classificação, foi utilizada uma biblioteca de aprendizagem de máquina denominada *scikit-learn*. A biblioteca, acessível para todos, é um conjunto de métodos simples e eficientes para mineração e analise de dados [2]. Também foi utilizada a biblioteca *matplotlib* para criar gráficos e a matriz de confusão [3]. E a biblioteca *scipy.io* para fazer a leitura dos dados [4].

III. Dados e Características

O banco de dados utilizado está disponível no site da Universidade de Stanford, do curso de redes neurais adaptativas. Os dados possuem 4.601 instancias de *e-mails* sendo 2.788 com *spam* e 1.813 sem, cada qual com 57 características o descrevendo.

No entanto, o banco de dados já possuía uma subdivisão entre casos de treinamento e de teste. Tal divisão não satisfazia nossos propósitos, e para ajusta-lo juntamos essa subdivisão e posteriormente a dividimos diversas vezes de forma aleatória em partições de 90% de treinamento para 10% de teste, assim podendo usar essas diversas divisões para aplicar o método de validação cruzada.

As 57 características descrevem a frequências das ocorrências de determinados termos e símbolos, selecionados pelo criador da base de dados, como possíveis indicadores de *spam*. Como por exemplo, os termos: grátis, dinheiro, crédito, negócios; e os símbolos: ';', '!', '$'.

IV. Proposta e Desenvolvimento

Com base no problema citado e nos dados que possuímos para analisarmos, buscaremos desenvolver metodologias para podermos classificar os *e-mails* com a menor taxa de erro possível.

Um grande cuidado que se deve ter na hora de classificar *e-mails* como *spam* ou não. É que não podemos permitir falsos positivos, assim um possível *e-mail* importante poderia

[1]Disponível em: https://www.python.org

[2]Disponível em: http://scikit-learn.org

[3]Disponível em: https://matplotlib.org

[4]Disponível em: https://docs.scipy.org/doc/

se perder como *spam*, sendo preferível então errar como falso negativo.

Com o propósito de descobrir a melhor abordagem, foram utilizado dois métodos de classificação: *Perceptron Multicamadas* e *Rede de Função de Base Radial* podendo assim realizar uma analise de desempenho e indicar os prós e contras de cada abordagem.

A. *Perceptron Multicamadas*

A Percepton Multicamadas é uma rede neural com mais de uma camada de neurônios em alimentação direta. Tal tipo de rede é composta por neurônios ligados entre si, onde suas ligações possuem pesos. Geralmente, neste tipo de rede neural, a aprendizagem é feita por retro-propagação do erro. A camada de entrada possui cada característica do que está sendo aprendido, e a camada de saída possui uma classificação do que está sendo analisado.

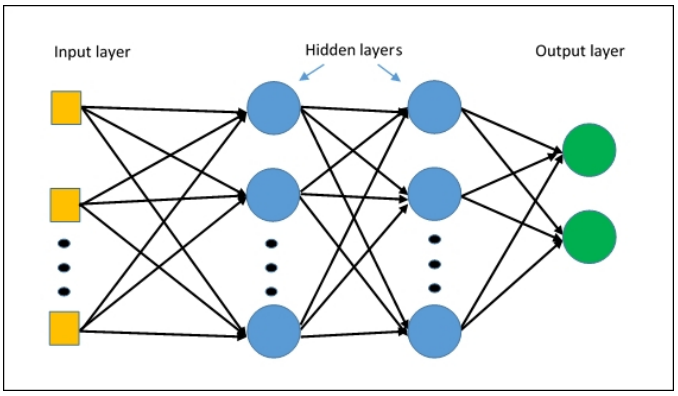

Figura 1. Diagrama esquemático de uma Rede de Perceptron Multicamadas

Para o problema proposto, usaremos na camada de entrada da rede Perceptron Multicamadas, como os 57 atributos indicadores de possibilidade de spam no *email*. E a saída é representada com um valor -1 ou 1 indicando a ausência ou presença de *spam* respectivamente.

B. *Rede de Função de Base Radial*

A rede neural com função de ativação de base radial consiste em um modelo neural multicamadas, capaz de aprender padrões complexos e resolver problemas não linearmente separáveis. A FBR contém três camadas, camada de entrada, na qual os padrões são apresentados à rede; a camada intermediária (única) que aplica a transformação não linear do espaço de entrada para o espaço escondido; e a camada de saída que fornece a resposta da rede ao padrão apresentado.

A principal diferença da FBR ao MLP é o fato de que a MLP utiliza hiperplanos para particionar o espaço de entrada e a FBR utiliza hiperelipsóides.

V. Resultados Experimentais

Como primeiro passo precisamos saber como os dados estão distribuídos. Pela Figura V podemos perceber que os dados seguem um padrão de distribuição do tipo F e possui uma grande quantidade de dados concentrados com características

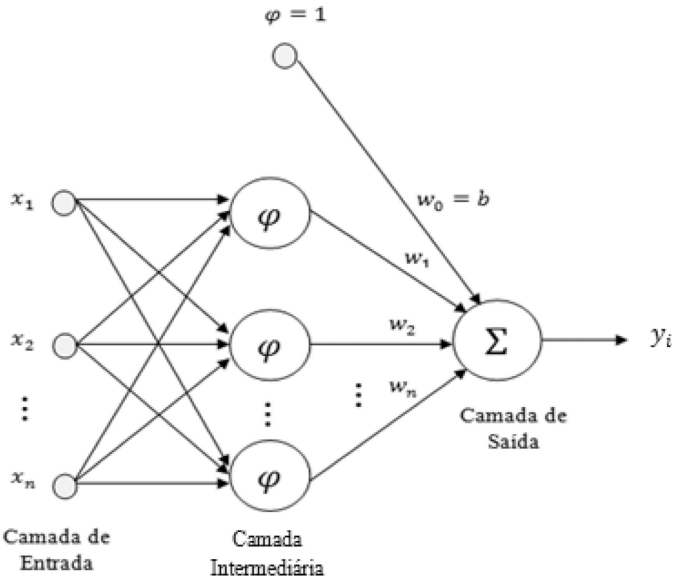

Figura 2. Diagrama esquemático de uma Rede de Função de Base Radial

próximas. Também podemos supor que esse fenômeno ocorre devido a troca de frequência das palavras, ou seja, se a palavra dinheiro aparece mais vezes isso implica em menos aparecimentos de outras palavras, o que leva a sempre equilibrar as frequências mantendo a média parecida para maioria.

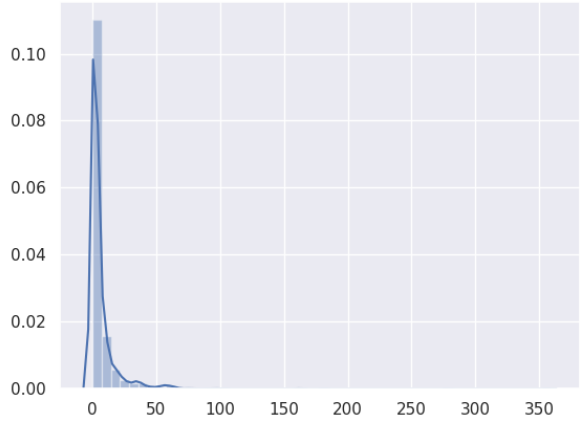

Figura 3. Distribuição da média dos dados

Para uma melhor noção da precisão adquirida testamos a Rede FBR com diferentes quantidade de centroide. Começamos com o valor de 10 em seguida 50, 100 e 200. Os resultados foram os seguintes:

- Para 10 centroides a média de precisão foi de 59%, a máxima de 67% e mínima de 54%.
- Para 50 centroides a média de precisão foi de 60%, a máxima de 63% e mínima de 58%. Aqui a média subiu um por cento mas os máximos e mínimos se concentraram mais mostrando uma menor variância.

- Para 100 centroides a média de precisão foi de 62%, a máxima de 66% e mínima de 60%.
- Para 200 centroides a média de precisão foi de 61%, a máxima de 66% e mínima de 56%.

A média não aumentar pode ser devido a densidade dos dados onde quando começamos a aumentar demais a quantidade de centroides resulta em uma grande intersecção dessas bases. Com isso cada nó pode classificar mesmos nós de maneira diferente por estarem agrupando dados diferentes.

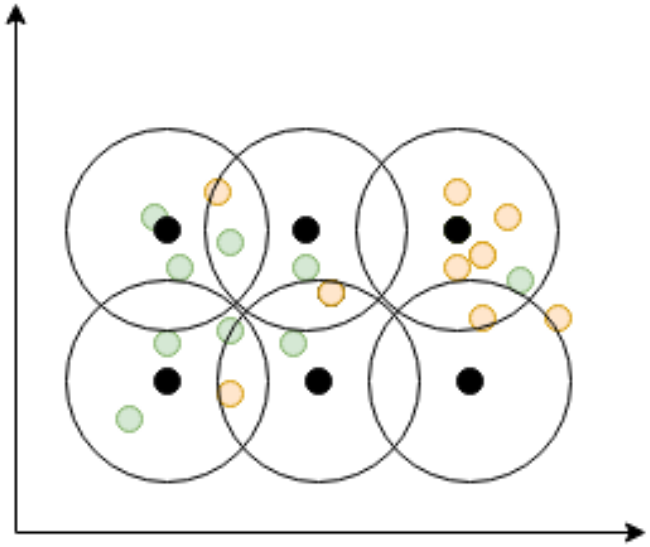

Figura 4. Exemplo de intersecção dos nós classificadores da FBR

Agora veremos os resultados expostos pela MLP. Com a biblioteca usada temos 3 diferentes algoritmos otimizados que podemos usar: *lbfgs*, *sgd* e *adam*.

Tabela I
TABELA DE ALGORITMOS MLP E SUAS PRECISÕES.

Algoritmo	Precisão
lsgbd	82%
adam	91%
sgd	60%

Experimentalmente o adam obteve o melhor resultado como visto na Tabela V. De acordo com a documentação do *scikit learn* o *lsgbd* pode possuir um melhor performance e velocidade com uma quantidade menor de dados, porém na maioria dos casos com dados em quantidade maior o *adam* se sai melhor.

Tendo em vista os resultados de ambas as redes neurais vemos que os dados são divididos melhor por uma função linear o que torna o MLP mais eficiente já que utiliza um hiperplano para fazer a classificação.

VI. CONCLUSÃO

Com os resultados mostrados podemos ter uma clara visão de o problema tende para uma resolução linear. Com esse trabalho vimos que as redes MLP são capazes de fazer classificações não lineares devido a sua possibilidade de

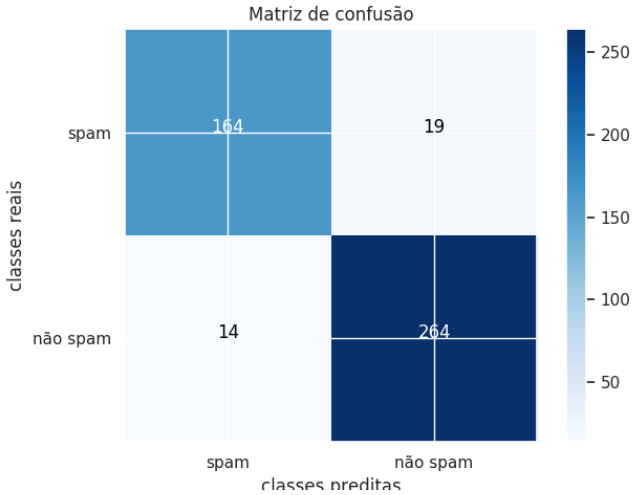

Figura 5. Matriz de Confusão MLP

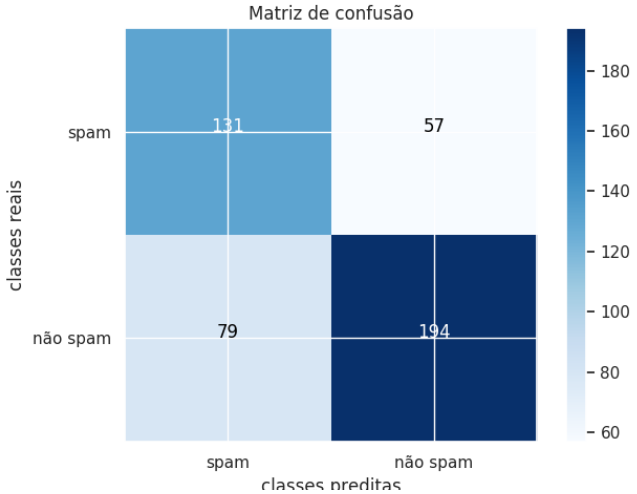

Figura 6. Matriz de Confusão FBR

múltiplas camadas e as FBRs necessitam de quantidade ótima de centroides para alcançar um bom rendimento. Além disso foi demonstrado a importância dos algoritmos de otimização disponíveis para MLPs sendo *adam* o com o melhor resultado.

Nota-se que os melhores resultados em relação a falsos positivos foram obtidos com o algoritmo do MLP. Tal resultado pode nos indicar uma melhor abordagem para a resolução do problema, visto que não queremos que o usuário tenha *emails* importantes considerados como *spam*. E o MLP também apresenta resultados melhores para falso negativos, diminuindo a quantidade de spam que o usuário recebe em sua caixa de *emails*.