Instituto Federal do Norte de Minas Gerais
campus Montes Claros
Curso de Bacharelado em Ciência da Computação

ESTUDO DE VIABILIDADE DE UM SISTEMA DE IDENTIFICAÇÃO DE CHAMADAS FALSAS EM CHAMADAS DE EMERGÊNCIA

ROBERT CRISTIANO ALMEIDA VIANA
Orientadora: Luciana Balieiro Cosme

Montes Claros
Abril de 2019

ROBERT CRISTIANO ALMEIDA VIANA

ESTUDO DE VIABILIDADE DE UM SISTEMA DE IDENTIFICAÇÃO DE CHAMADAS FALSAS EM CHAMADAS DE EMERGÊNCIA

Projeto de Monografia apresentado ao Curso de Graduação em Ciência da Computação do Instituto Federal do Norte de Minas Gerais – Campus Montes Claros, como requisito parcial para a obtenção do grau de Bacharel em Ciência da Computação.

Orientadora: Luciana Balieiro Cosme

Montes Claros

Abril de 2019

"Se não fosse difícil, todo mundo faria isso. É a dificuldade que torna isso grandioso"
(Tom Hanks)

Resumo

Palavras-chave: Inteligência Artificial, Classificação de texto, Redes Neurais, Processamento de Linguagem Natural.

Abstract

Keywords: Artificial Intelligence.

Sumário

Lista de Figuras

Capítulo 1

Introdução

1.1 Motivação

1.2 Objetivos

1.3 Estrutura do trabalho

Capítulo 2

Conceitos Básicos

Nesta seção, iremos explorar alguns fundamentos básicos para compreender, passo a passo, como será conduzida a metodologia proposta. Primeiramente será introduzido uma abstração matemática muito utilizada na computação, o Grafo, que têm como serventia facilitar a visualização de como é composto uma estrutura de um neurônio artificial. Feito isso, descrevemos brevemente sobre classificação de dados e também sobre duas técnicas de categorização, as Redes Neurais Artificiais (RNA) e o K-vizinhos mais próximos (do inglês: *K-Nearest Neighbors* - KNN). Por fim, é apresentado uma breve introdução ao Processamento de Linguagem Natural (PLN), uma sub-área da da Inteligência Artificial (IA).

2.1 Grafos

Muitas situações no mundo real podem ser descritas com o uso de um diagrama, composto por um conjunto de pontos e arestas, onde as arestas unem pares desses pontos. Por exemplo, os pontos podem representar cidades em um mapa, e as arestas representariam as estradas que ligam duas cidades. O conceito de grafo parte de uma abstração matemática para caracterizar situações com essas características (Bondy & Murty, 1976).

Matematicamente um grafo G é uma tripla (V, E, ψ), consistido por um conjunto não vazio de vértices V, um conjunto de arestas E e uma função de incidência ψ que caracteriza quais vértices possuem uma relação (através de uma aresta) com outros vértices. Por exemplo, seja $G = (V, E, \psi)$ um grafo (Figura 2.1), tal que $V = \{0, 1, 2, 3, 4, 5\}$, $E = \{a, b, c, d, e, f, g, i\}$ e ψ a função incidência representada na Tabela 2.1.

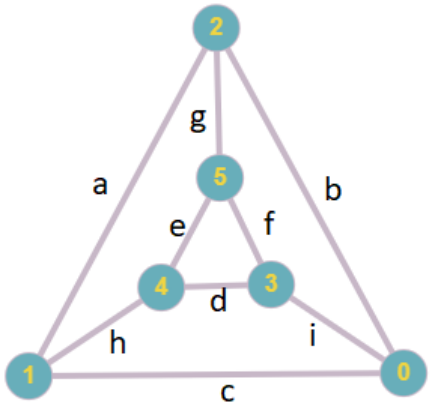
Segundo Bondy & Murty (1976), os grafos possuem esse nome porque eles possuem uma representação gráfica, e são essas representações que facilitam o entendimento de suas propriedades.

Tabela 2.1. Função incidência ψ de G

$\psi_a = 1, 2$
$\psi_b = 2, 0$
$\psi_c = 1, 0$
$\psi_d = 4, 3$
$\psi_e = 4, 5$
$\psi_f = 5, 3$
$\psi_g = 2, 5$
$\psi_h = 1, 4$
$\psi_i = 4, 3$

Fonte: Elaborado pelo autor

Figura 2.1. Diagrama do Grafo G



Fonte: Graph Online (2015).

2.2 Classificação de dados

Classificação de dados é um problema que abrange enumeras aplicações em diversos tipos de cenários no nosso dia a dia, tais como diagnóstico de doenças, identificação de objetos em fotos e vídeos, categorização de seres vivos e espécies, dentre outros. Esse problema é um dos tópicos mais ativos na área de aprendizado de máquina. Isso se dá, porque classificar dados consiste em determinar um rótulo ou classe para um objeto, baseado em um conjunto de características extraídas do mesmo (Duda et al., 1973; Bishop, 2006).

Em geral, dados são classificados como pertencentes a uma única classe ou categoria. Essa forma de classificação é denominada classificação de rótulo único. Por outro lado, se houver mais uma forma de rotular a mesma entrada, então dá-se o nome de classificação de multi-rótulo.

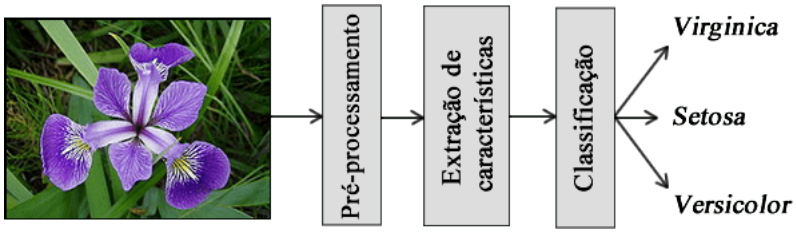
Formalmente o processo de classificação consiste em $X = \{X_1, ..., X_i\}$ um conjunto de i entradas, $C = \{c_1, ..., c_n\}$ um conjunto de n classes, tal que $n \geq 2$, e

$Y = \{(X_1, \{c_1, ..., c_j\}), ..., (X_i, \{c_n, ..., c_k\})\}$ um conjunto de treinamento, no qual cada entrada X_i é categorizada por uma ou mais classes c_i. O objetivo geral de um classificador é aprender, através de seu conjunto de treinamento Y, uma possível correlação entre os atributos das entradas com suas classes, de tal forma que para uma entrada $X' = \{X'_1, ..., X'_i\}$ que não possua rótulo c qualquer, seja possível classificar-lá.

Para ilustrar o processo de classificação de dados, considere o problema da flor de Iris. Nesse problema, existe um conjunto de flores do gênero Iris que podem ser rotuladas de uma das três maneiras: do tipo setosa, virgínica ou versicolor. Partindo desse ponto, o objetivo é determinar a qual grupo uma determinada flor pertence baseado nas medidas de sépalas e pétalas da mesma. A Figura 2.2 ilustra o processo de classificação. Inicialmente as informações especificas sobre as sépalas e pétalas devem ser extraídas em um pré-processamento. Em seguida tais medidas são processadas e suas características extraídas. Por fim, é realizada a classificação das flores. Neste exemplo os valores de X serão as medidas de comprimento, largura das sépalas e pétalas e C assumirá os rótulos setosa, virgínica e versicolor.

Em geral, existem diversos algoritmos para classificação de dados, onde cada um possui sua especificidade, vantagens e desvantagens. Neste trabalho, aborda-se o uso de duas técnicas para classificação de dados, sendo elas RNA e KNN.

Figura 2.2. Processo de classificação de flores do gênero Iris



Fonte: Pacheco (2016)

2.2.1 Redes neurais artificiais

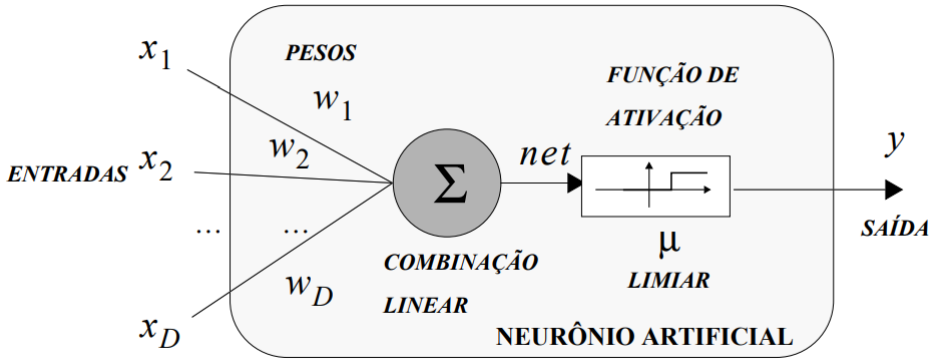
O ser humano possui capacidades cognitivas extraordinárias e, desde o surgimento da computação, desejou-se projetar máquinas capazes de realizar tarefas inteligentes que, até então, somente eram executadas por humanos. Os primeiros trabalhos desenvolvidos nessa área foram: um neurônio apresentado por McCulloch & Pitts (1943), usado posteriormente como base para a concepção do *Perceptron* por Rosenblatt (1958) e um neurônio chamado *Adaline* por Widrow & Hoff (1960). Tais trabalhos deram origem ao conceito de RNA que, em outras palavras, é uma tentativa de copiar a estrutura e o funcionamento do cérebro, composto este por bilhões de neurônios, para uma estrutura

artificial, transformando assim as redes neurais biológicas em redes neurais artificiais (Rauber, 2005).

Para compreender o conceito por trás de uma rede neural, é preciso introduzir um modelo simplificado de um neurônio e suas capacidades de processamento associadas. Cada neurônio é considerado como uma unidade básica de processamento que, quando estimulada por sinais de entrada, emitem sinais de saída como uma reação. Tais sinais emitidos por um neurônio, são repassados para outros neurônios através de uma conexão sináptica. Tal processo pode ser repetido por várias camadas de neurônios até chegar ao nosso cérebro, que então processa essa informação e produz novas reações (Baeza-Yates et al., 1999). A principal função de uma rede neural é armazenar conhecimento experimental e torná-lo disponível, o que em prática significa que este conhecimento é adquirido e armazenado em pesos sinápticos durante o processo. Uma RNA é normalmente implementada através de um programa de computador (*software*) ou através de componentes eletrônicos (*hardware*).

Uma rede neural pode ser representada matematicamente através de uma estrutura de grafo (Figura 2.3), em que os vértices fazem o papel dos neurônios e as arestas representam as conexões sinápticas entre os neurônios, no qual se adicionarmos pesos a tais arestas, é possível mensurar a força de tal conexão sináptica. Seja x_i entradas fornecidas por outros neurônios para um neurônio artificial. O processamento desse neurônio consiste em uma combinação linear das D entradas tais que $\sum_{i=1}^{D} = w_i x_i$, onde x_i é uma aresta com peso w_i. Se tal valor ultrapassar um limiar μ, esse neurônio dispara um valor positivo (1) na saída binária y, caso contrário dispara um valor negativo (0) na saída.

Figura 2.3. Diagrama de um neurônio artificial



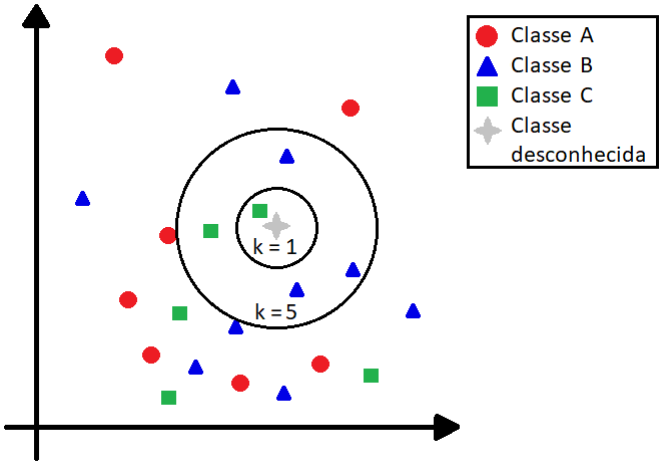
Fonte: Rauber (2005)

2.2.2 K-vizinhos mais próximos

O algoritmo K-vizinhos mais próximos (do inglês: *K-neareast neighbours* - KNN) tem como objetivo determinar o rótulo de classificação de uma amostra, baseando-se em outras amostras vizinhas, advindas de um conjunto de treinamento. O classificador KNN, um dois mais simples algoritmos de classificação, é baseado em instâncias. Esse algoritmo encontra os k objetos mais similares e realiza uma votação de acordo com as classes às quais pertencem esses k objetos, assinalando por fim, uma classe ao objeto de teste. A literatura apresenta diversas formas para expressar essa distância/similaridade dentre os objetos de análise (Fukunaga & Narendra, 1975; Duda et al., 1973). Por exemplo, se os dados trabalhados estão em formato de texto, é comum utilizar a similaridade por cossenos. Por outro lado, se os dados possuírem formato numérico, possivelmente a distância euclidiana será mais eficaz.

Na Figura 2.4, é ilustrado o processo de classificação com o algoritmo KNN. Neste exemplo, têm-se três classes, anteriormente conhecidas, sendo elas: classe A (círculo vermelho), classe B (triangulo azul) e classe C (quadrado verde). O objetivo é identificar, por similaridade, a qual classe pertence a amostra (estrela cinza), olhando para o seus k vizinhos mais próximos. Para $k = 1$, esse algoritmo classificaria a amostra como pertencente a classe C. Por outro lado, se o valor escolhido para k é 5, por votação majoritária, a amostra seria classificada como pertencente a classe B.

Figura 2.4. Exemplo do K-vizinhos mais próximos em um plano cartesiano



Fonte: Elaborado pelo autor

2.3 Processamento de linguagem natural

O processamento de linguagem natural (PLN) têm como objetivo tratar os mais diversos aspectos presentes dentro da comunicação humana, tais como sons, palavras, sentenças e discursos, levando em consideração os seus formatos, referências, estruturas, significados, contextos e aplicações. Embora exista outros animais que possuem um vocabulário com centenas de sinais, tais como os elefantes e os golfinhos, somente os seres humanos possuem a capacidade de se comunicar, de forma confiável, em um número ilimitado de mensagens qualitativamente diferentes, sobre um tema qualquer (Russell & Norvig, 1994; Gonzalez & Lima, 2003).

Hoje em dia, com o constante crescimento da rede mundial de computadores, possibilitou o acesso a enumeras páginas de informações na *Web*, no qual quase todas elas estão em um formato de linguagem natural. Entretanto, disponibilidade não significa fácil acesso à informação. Para uma máquina adquirir tal conhecimento, ela precisa ser treinada, de forma exaustiva, para compreender as complexas, e muitas vezes ambíguas, linguagens em que os seres humanos se comunicam.

Segundo Russell & Norvig (1994) as linguagens naturais, tais como o português e o espanhol, não podem ser caracterizadas como um conjunto de sentenças definitivas, pois de acordo com o contexto em que for definida uma sentença de alguma dessas linguagens, ela pode possuir inúmeras interpretações diferentes. Portanto, convém definir um modelo de linguagem natural como uma distribuição de probabilidade sobre sentenças. Existe um famoso ditado popular brasileiro que diz "para um bom entendedor, meia palavra basta", o que pode ser comumente aplicado para nós humanos, que possuímos uma espécie de dispositivo de especialização para aquisição de linguagens (Chomsky, 2014). Já que meia palavra basta, pode-se concluir que uma sentença de uma linguagem natural não é sempre aleatória, e que sim possui algum grau de previsibilidade e correlação entre a escolha das palavras. Portanto, nos leva a acreditar que palavras similares estejam presentes no mesmo contexto.

O PLN consiste no emprego de um conjunto de técnicas computacionais para aprender, entender e reproduzir uma linguagem natural. No processo de tradução do significado, tratamento de ambiguidade e entre outros desafios, o PLN pode utilizar de conhecimentos linguísticos e métodos estatísticos para resolver-los. Por exemplo, considere uma análise sobre dois textos semelhantes A e B, no qual desconfia-se que exista possibilidade de plágio. Com o uso de um pré-processamento, seria possível filtrar os textos para remover *Stopwords*, que são palavras funcionais, tais como artigos, preposições e conetivos, que quando analisadas individualmente não possuem grande relevância para o contexto. Após o pré-processamento, é possível aplicar um método da distância mínima de edição, que como o próprio nome diz, significa quantas operações de

inserções, remoções ou substituições de caracteres são necessárias para que o texto A torne-se o texto B, ou vice-versa. Para compreender melhor o processo de aplicação de conhecimentos linguísticos, é apresentado a seguir uma seção sobre classificação de texto.

2.3.1 Classificação de texto

Segundo Aggarwal (2014) um dos principais desafios encontrados durante o processo de classificação de texto é sobre o tamanho dos dados tratados, que podem variar de algumas poucas dezenas para milhões de palavras. Esses dados se encontram, quase sempre, de maneira esparsa, ou seja, possuindo baixa frequência de uso. Por outro lado, têm-se muitas vezes uma alta frequência de dados não úteis para tratamento, como as *Stopwords*. Por esse motivo, uma etapa de extração e seleção de características é essencial, e pode gerar benefícios como redução da dimensão do problema, que por sua vez aumenta a velocidade de execução do algoritmo; redução na quantidade total de características; aumento na precisão de predição e acerto; e facilitar a visualização dos dados.

Uma vez escolhido um conjunto de características, é possível aplicar algumas técnicas de classificação de texto, inclusive duas das que já foram abordadas aqui, as redes neurais artificiais e o k-vizinhos mais próximos, nos quais apresentam precisão superior a 98% quando aplicados ao problema de identificação de *e-mail spam* (Russell & Norvig, 1994).

A classificação ou categorização de texto é a tarefa de, dado algum tipo de texto, decidir a qual conjunto predefinido de classes o mesmo pertence. Por exemplo, decidir a qual linguagem pertence uma sentença ou identificar se um texto de uma chamada de emergência é falsa ou não, são exemplos de classificação de texto.

Capítulo 3

Referencial Teórico

Capítulo 4

Metodologia

4.1 Cronograma

Referências Bibliográficas

Aggarwal, C. C. (2014). *Data classification: algorithms and applications*. CRC press.

Baeza-Yates, R.; Ribeiro-Neto, B. et al. (1999). *Modern information retrieval*, volume 463. ACM press New York.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Bondy, J. A. & Murty, U. S. R. (1976). *Graph Theory with Applications*. Elsevier, New York.

Chomsky, N. (2014). *Aspects of the Theory of Syntax*, volume 11. MIT press.

Duda, R. O.; Hart, P. E. et al. (1973). *Pattern classification and scene analysis*, volume 3. Wiley New York.

Fukunaga, K. & Narendra, P. M. (1975). A branch and bound algorithm for computing k-nearest neighbors. *IEEE Transactions on Computers*, C-24(7):750–753.

Gonzalez, M. & Lima, V. L. (2003). Recuperação de informação e processamento da linguagem natural. In *XXIII Congresso da Sociedade Brasileira de Computaçao*, volume 3, pp. 347–395.

Graph Online (2015). An online project aimed at creation and easy visualization of graph and shortest path searching. Disponível em https://graphonline.ru/en/?graph=Cayleygraph. Acesso em 13 de Maio de 2019.

McCulloch, W. S. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

Pacheco, A. G. C. (2016). Agregação de classificadores neurais via integral de choquet com respeito a uma medida fuzzy. Master's thesis, Programa de Pós-Graduação em Informática.

Rauber, T. W. (2005). Redes neurais artificiais. *Universidade Federal do Espírito Santo*.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

Russell, S. J. & Norvig, P. (1994). *Inteligência artificial*. Pearson Educación.

Widrow, B. & Hoff, M. E. (1960). Adaptive switching circuits. Technical report, Stanford Univ Ca Stanford Electronics Labs.