Avaliação de medidas de rede complexas para classificação de dados via caracterização de importância

Guilherme Massahiro Suzuki

Universidade Federal de Uberlândia
Faculdade de Computação
Bacharelado em Sistemas de Informação

Monte Carmelo - MG
2023

Guilherme Massahiro Suzuki

Avaliação de medidas de rede complexas para classificação de dados via caracterização de importância

Trabalho de Conclusão de Curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, Minas Gerais, como requisito exigido parcial à obtenção do grau de Bacharel em Sistemas de Informação.

Área de concentração: Sistemas de Informação

Orientador: Prof. Dr. Murillo Guimarães Carneiro

Monte Carmelo - MG
2023

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Faculdade de Computação
Av. João Naves de Ávila, nº 2121, Bloco 1A - Bairro Santa Mônica, Uberlândia-MG, CEP 38400-902
Telefone: (34) 3239-4144 - http://www.portal.facom.ufu.br/ facom@ufu.br

ATA DE DEFESA - GRADUAÇÃO

Curso de Graduação em:	Bacharelado em Sistemas de Informação				
Defesa de:	Trabalho de Conclusão de Curso II				
Data:	30/11/2023	Hora de início:	08:00	Hora de encerramento:	09:10
Matrícula do Discente:	31611BSI001				
Nome do Discente:	Guilherme Massahiro Suzuki				
Título do Trabalho:	Avaliação de medidas de rede complexas para classificação de dados via caracterização de importância				
A carga horária curricular foi cumprida integralmente?	(X) Sim ( ) Não				

Reuniu-se na Sala 1B-230, Campus Santa Mônica, da Universidade Federal de Uberlândia, a Banca Examinadora, designada pelo Colegiado do Curso de Graduação em Bacharelado em Sistemas de Informação - Campus Monte Carmelo, assim composta: Professores: Dra. Maria Adriana Vidigal de Lima (FACOM/UFU), Dr. Marcelo Zanchetta do Nascimento (FACOM/UFU) e Dr. Murillo Guimarães Carneiro (FACOM/UFU), orientador do candidato.

Iniciando os trabalhos, o presidente da mesa, Dr. Murillo Guimarães Carneiro, apresentou a Comissão Examinadora e o candidato, agradeceu a presença do público, e concedeu ao discente a palavra, para a exposição do seu trabalho. A duração da apresentação do discente e o tempo de arguição e resposta foram conforme as normas do curso.

A seguir o senhor presidente concedeu a palavra, pela ordem sucessivamente, aos examinadores, que passaram a arguir o candidato. Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando o candidato:

(X) Aprovado(a) Nota [95]

OU

( ) Aprovado(a) sem nota.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme foi assinada pela Banca Examinadora.

Documento assinado eletronicamente por **Murillo Guimarães Carneiro**, **Professor(a) do Magistério Superior**, em 01/12/2023, às 10:12, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Marcelo Zanchetta do Nascimento**, **Professor(a) do Magistério Superior**, em 01/12/2023, às 10:17, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Maria Adriana Vidigal de Lima**, **Professor(a) do Magistério Superior**, em 01/12/2023, às 10:49, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5008350** e o código CRC **61B0A547**.

Referência: Processo nº 23117.017360/2021-71 SEI nº 5008350

Dedico a minha familia por me darem a oportunidade de me dedicar a minha formação e pelo apoio que sempre me deram para que isso fosse possível, e a todos aqueles que acreditaram e fizeram parte dessa trajetória, e em especial ao Professor e amigo Dr. Murillo Carneiro por todo apoio, tempo e dedicação para que esse trabalho fosse concluído.

Agradecimentos

Ao Dr. Daisaku Ikeda por todo ensinamento que pôde me dar, por me guiar e mostrar que tudo depende de mim e que não posso medir esforços para conquistar tudo aquilo que almejo.

A minha mãe Vania por não medir esforços para me tornar a pessoa que sou hoje, por sempre estar do meu lado quando eu precisei e por toda a coragem que demonstrou ter nos momentos mais difíceis.

Ao meu pai Masayuki por sempre ter me mostrado o exemplo a ser seguido, me motivando e celebrando comigo todas as conquistas.

A minha irmã Patricia por sempre estar me apoiando nos momentos mais difíceis, e me impulsionando para que eu sempre me desenvolva e me torne sempre melhor e que tem muita importância no meu crescimento e desenvolvimento até hoje.

A minha namorada Karolaine por sempre estar do meu lado me ensinando e dando todo o suporte necessário.

A professora Daniele Carvalho Oliveira que me desenvolveu fortemente durante meus estudos, por toda dedicação e seriedade em planejar e treinar as equipes de maratonas em que pude participar e por todo apoio dentro e fora da universidade.

Ao professor Dr. Murillo Carneiro, que me acompanhou nessa trajetória de trabalhos e pesquisa, que me incentivou fortemente para a defesa deste trabalho, pela dedicação e ética com o seu trabalho como professor, orientador e amigo, por todo ensinamento como professor e orientador.

A todos aqueles que tive contato durante os meus estudos na universidade, em especial aqueles que pude levar para fora da universidade que tornaram grandes amigos.

"Não faz mal que seja pouco, o que importa é que o avanço de hoje seja maior que o de ontem, que nossos passos de amanhã sejam mais largos que os de hoje. Atuem agora e vivam o presente com a certeza de que neste exato instante está se erguendo o futuro. Deixem seus méritos gravados na história de suas contínuas vitórias! A dificuldade no momento presente será a glória em seu futuro."

(Dr. Daisaku Ikeda)

Resumo

O uso de redes complexas se tornou um tópico cada vez mais falado em vários domínios incluindo a classificação de dados. Por utilizarem apenas características físicas dos dados classificadores tradicionais possuem dificuldade em classificar um novo dado baseado em seu padrão estrutural ou topológico, tal como fazem técnicas de classificação baseadas em redes complexas para caracterização de importância ou conformidade de padrão. Este trabalho consiste em investigar o modelo de classificação via caracterização de importância a fim de analisar empiricamente outras medidas de redes complexas em comparação a medida *PageRank*. O ambiente desenvolvido para realização destes testes e posteriormente das análises é feito a partir de um grafo contruído a partir da rede k-vizinhos mais próximos, e a utilização de medidas de redes complexas a capazes de caracterizar algum tipo de importância entre os vértices da rede. Para isso foram selecionadas 4 medidas de redes complexas, a saber: grau, proximidade, *PageRank* e menor caminho médio. Para os experimentos foi utilizadas 2 bases artificiais com diferentes níveis de ruídos e outras 4 bases reais. Os resultados obtidos demonstram que outras medidas de redes podem ter o resultado competitivo em comparação ao PageRank, e que algumas medidas de redes podem ser bem sensíveis em relação a etapa de construção da rede. De forma concisa, este trabalho tem grande potencial para auxiliar trabalhos relacionados à classificação de dados via caracterização de importância bem como a seleção mais adequada de medidas de redes complexas para tarefas de classificação.

Palavras-chave: Redes complexas, Classificação de dados, Classificação baseada em importância, Classificação de alto nível, Conformidade de padrão.

Abstract

The use of complex networks has gained attention in various domains including data classification. Because they essentially consider physical characteristics of data, traditional classifiers have difficulty classifying new data based on its structural or topological pattern, as do classification techniques based on complex networks like importance characterization or pattern conformity. This work consists of investigating the classification model via importance characterization in order to empirically analyze other measures of complex networks in comparison to the PageRank measure. The environment developed to carry out these tests and subsequently the analyses is made from a graph constructed using the k-nearest neighbors network, and the use of measures of complex networks capable of characterizing some type of importance between the vertices of the network. For this purpose, 4 measures of complex networks were selected, namely: degree, closeness, PageRank and average shortest path. For the experiments we used 2 artificial datasets with varying levels of noise and another 4 real-world datasets. The obatined results demonstrate that other network measures can have a competitive result compared to PageRank, and that some network measures can be very sensitive in relation to the network construction phase. Concisely, this work has potential to assist other works related to data classification via importance characterization or even those regarding the appropriate selection of complex network measures for classification tasks.

Keywords: Complex networks, data classification, classification by data importance, classification by pattern conformation, high level classification.

Lista de ilustrações

Lista de tabelas

Sumário

Capítulo 1

Introdução

A classificação é uma das mais conhecidas tarefas do aprendizado de máquina, e possui bastante associação com tarefas cognitivas humanas. De maneira natural, o ser humano sempre está classificando algo ao seu redor: classificando pessoas no seu círculo de amizades (colegas de trabalho, familiares, amigos, colegas de faculdade, etc.) ou comidas (salgado, doce, entrada, prato principal, sobremesa, etc.). Tais classificações são frutos de diferentes situações e contextos e muitas das vezes não estão apenas relacionadas à semelhança física entre objetos, mas ao conceito semântico que se deseja representar. Por exemplo, ao nos depararmos com uma mesa, seja lá qual for seu material, tipo, formato, ainda assim sabemos que é uma mesa e a classificamos como tal (CARNEIRO; ZHAO, 2018).

Do ponto de vista da computação, a classificação de dados consiste em aprender um modelo ou função, denominado classificador, baseada em um conjunto de dados já conhecidos para classificar um novo dado em uma das classes. Exemplos de técnicas convencionais de classificação incluem árvore de decisão, aprendizado baseado em instâncias, redes neurais e máquina de vetores de suporte (CARNEIRO; ZHAO, 2017). Em comum, tais técnicas realizam a classificação considerando apenas as características físicas (distância ou distribuição), o que significa que elas podem ter dificuldades para capturar propriedades semânticas dos dados, tais como a formação de padrão (RESENDE; CARNEIRO, 2021). Outro problema é que a maioria dessas técnicas julgam que todos os objetos possuem a mesma relevância, no entanto, negligenciar a relevância individual de cada objeto pode alterar a compreensão do problema (CARNEIRO; ZHAO, 2018).

Ao utilizar redes complexas, elas podem nos prover diferentes conceitos e heurísticas para a tarefa de classificação, por considerar não só as características físicas dos dados, mas também a sua estrutura semântica (CARNEIRO; ZHAO, 2017). Dessa forma, a partir da representação de relações funcionais, espaciais e topológicas em redes, é possível capturar, por exemplo, a formação de padrão nos dados (CARNEIRO et al., 2019).

A Figura 1 representa um exemplo da classificação de técnicas convencionais. Demonstra que o grupo de itens de teste que estão representados como triângulos pretos formam

um padrão muito mais parecido com a classe dos círculos azuis do que com o grupo dos quadrados vermelhos. Erroneamente, classificadores convencionais os classificam como sendo da classe dos quadrados vermelhos, pois não conseguem capturar a formação de padrões dos itens.

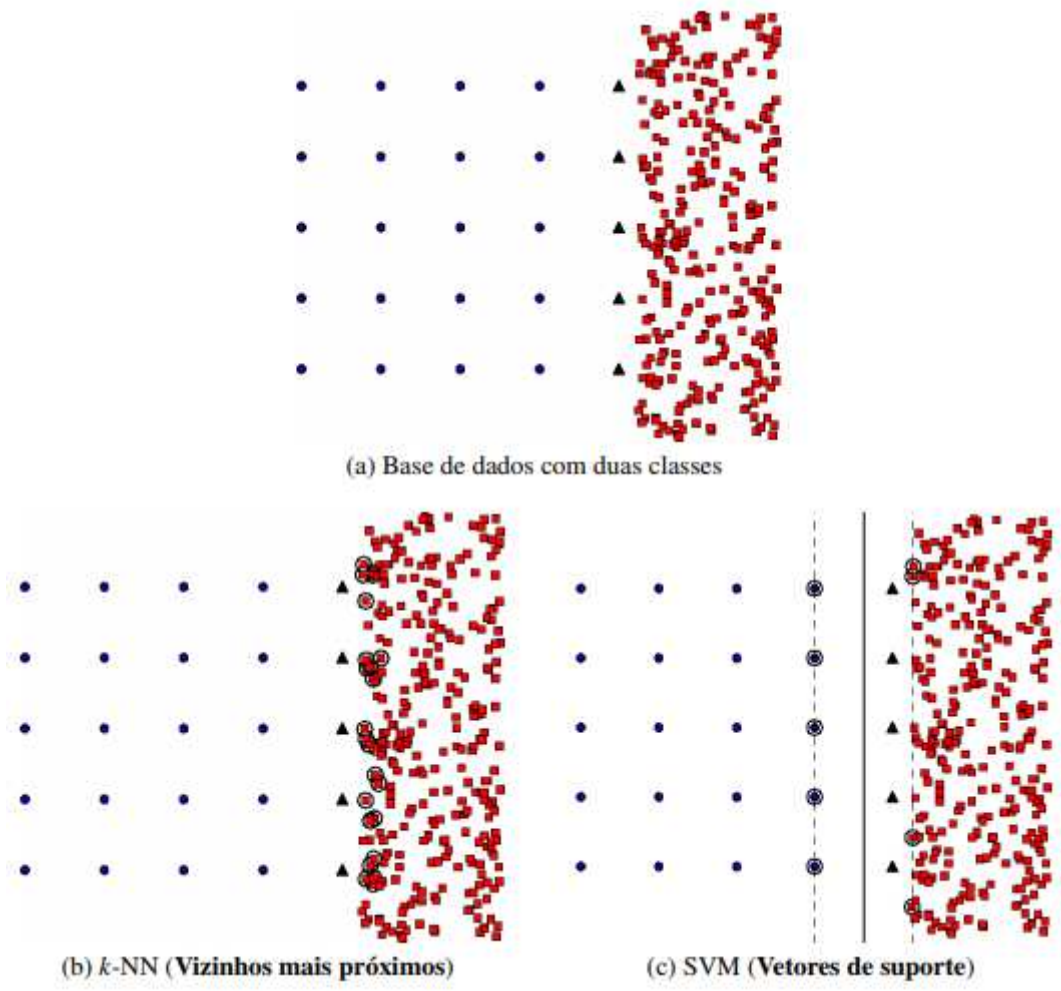


Figura 1 – Processo de classificação de técnicas convencionais. (a) Base de dados artificial, apresenta duas classes círculos azuis e quadrados vermelhos, itens de teste são representados por triangulos pretos. (b) k-vizinhos mais próximos classifica os itens de teste como quadrado vermelho. (c) Máquina de vetores de suporte também rotula como pertencente a classe quadrado vermelho. Técnicas convencionais são incapazes de considerar a estrutura semântica dos dados. Retirado de (CARNEIRO; ZHAO, 2017)

A classificação de dados via caracterização de importância é uma técnica proposta recentemente capaz de considerar tanto características físicas quanto estruturais dos dados de entrada. Nessa técnica, cada objeto de treinamento recebe um valor de importância e quando se é introduzido um novo objeto de teste, ele é classificado na classe em que ele recebe maior importância (CARNEIRO; ZHAO, 2018). O cálculo da importância é derivada do algoritmo *PageRank*, o mesmo algoritmo usado no motor de busca da Google. Nesse algoritmo, quanto mais vértices ligados a um objeto, maior é a importância dele,

e se um objeto recebe ligações de vértices importantes, isso quer dizer que ele também é considerado importante. Alguns trabalhos que mostraram a importância de se obter informações estruturais e topológicas estão apresentados em (CARNEIRO et al., 2023) e (CARNEIRO; ZHAO, 2018), que serão explorados com mais detalhes na subseção 2.3.4.

1.1 Motivação

As motivações para que sejam avaliadas as técnicas utilizadas para a classificação de dados utilizando redes complexas, estão relacionadas às diferentes heurísticas de classificação que o aprendizado baseado em redes pode oferecer. Dentro da literatura há estudos de técnicas como redes neurais, árvores de decisão, máquina de vetores de suporte. O que todas elas têm em comum é que para realizar a tarefa de classificação, consideram apenas os atributos físicos dos dados, o que torna um limitador no sentido de que os dados possam estar distribuídos de forma que características topológicas, funcionais e espaciais se tornem essenciais para uma classificação ótima.

Estudos recentes mostram que há maneiras de combinar técnicas de diferentes níveis para a tarefa de classificação como a proposta em (CARNEIRO; ZHAO, 2018), baseada em conceitos de redes complexas. Apesar dos aspectos salientes alcançados por aquela técnica, naquele trabalho os autores avaliaram apenas o *PageRank* enquanto medida para caracterizar a importância dos dados. Dessa forma, a investigação aqui proposta pretende estendê-lo por considerar e avaliar várias outras medidas de rede disponíveis na literatura.

1.2 Objetivos e Desafios da Pesquisa

O principal objetivo deste trabalho é desenvolver novas heurísticas baseadas em teorias de redes complexas para a classificação de dados via caracterização de importância. Especificamente, vislumbram-se os seguintes objetivos:

- ❏ Reproduzir a técnica de classificação baseada em importância proposta em Carneiro e Zhao (2018).
- ❏ Analisar as medidas de redes complexas mais conhecidas na literatura e sua aplicação no contexto da classificação de dados via caracterização de importância.
- ❏ Avaliar empiricamente novos métodos de classificação de dados baseado em importância a partir de outras medidas além do *PageRank*.

1.3 Hipótese

O uso de outras medidas de caracterização de importância baseada em redes complexas além do *PageRank* pode prover novas heurísticas eficientes para a classificação de dados via caracterização de importância.

1.4 Organização da Monografia

O restante deste trabalho está organizado da seguinte forma:

No capítulo 2 são apresentados conceitos e fundamentos teóricos para a compreensão do trabalho proposto: classificação de dados, redes complexas, medidas de redes, aprendizado de máquinas em redes complexas.

O capítulo 3 expõe todos os materiais e métodos utilizados para a realização das simulações, as bases de dados reais e artificiais, a técnica de classificação de dados via importância, medidas de redes complexas e o ambiente experimental.

O capítulo 4 discute os resultados obtidos pelos testes e as comparações com a medida *PageRank*, assim como a sensibilidade das medidas de redes em relação a variação dos parâmetros.

No capítulo 5 são relatadas as conclusões baseadas nos experimentos feitos e resultados obtidos e próximos passos para trabalhos futuros, bem como a produção bibliográfica relacionada à monografia.

Capítulo 2

Fundamentação Teórica

Neste capítulo serão apresentados fundamentos de classificação de dados, redes complexas e classificação de dados baseada em redes complexas. Especificamente, serão discutidos a classificação de dados, assim como as etapas para a classificação e conceitos e medidas de redes complexas. Também serão discutidos alguns algoritmos e técnicas de aprendizado de máquina relacionados, que utilizam conceitos de redes complexas.

2.1 Classificação de dados

A classificação de dados é uma tarefa de aprendizado de máquina cujo objetivo é aprender uma função alvo F que mapeie cada conjunto de atributos x para um dos rótulos de classes y pré-determinados (TAN; STEINBACH; KUMAR, 2009). Para isso é necessário treinar um modelo de classificação com um conjunto de dados contendo instâncias em que o rótulo de classe já é conhecido. Muitas das situações do mundo real podem ser modeladas como problemas de classificação, por exemplo, a classificação de um novo *e-mail* como "*spam*" ou "não *spam*" (AGGARWAL, 2014).

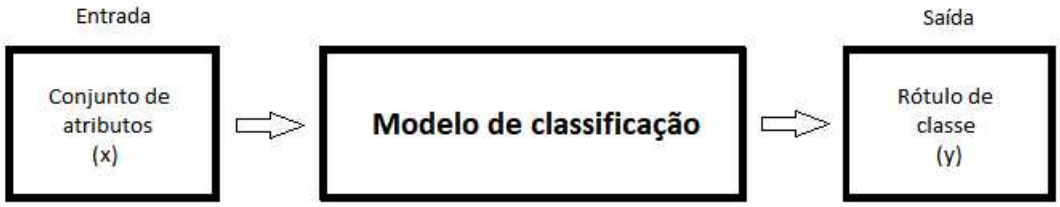


Figura 2 – Classificação como a tarefa de mapear um conjunto de atributos x no seu rótulo de classe y (TAN; STEINBACH; KUMAR, 2009).

Considerando o problema de detectar se um *e-mail* é *spam* ou não, o aprendizado de máquina tem como objetivo criar um modelo de classificação que aprenda a partir de um conjunto de mensagens de um dado domínio, algumas delas classificadas como *spam* ou

não por um especialista, classificar novas mensagens (cuja designação de *spam* ou não *spam* é desconhecida) através de tal modelo (CARNEIRO; ZHAO, 2017).

É possível dividir o aprendizado de máquina em três tipos de aprendizado: supervisionado, não supervisionado e o semissupervisionado. A principal diferença entre o aprendizado não supervisionado em relação aos outros é a ausência de informação de classes. Isso torna difícil quantificar a performance do modelo. Agrupamento e redução de dimensionalidade são exemplos de aprendizado não supervisionado.

No aprendizado supervisionado o objetivo é conseguir classificar a partir de uma instância de entrada a saída desejada correspondente. A principal característica é a presença de um supervisor, o qual determina os rótulos corretos para cada entrada dos dados de treinamento. A classificação, regressão e o ranqueamento são exemplos (MOHRI; ROSTAMIZADEH; TALWALKAR, 2018). O problema de *spam* comentado anteriormente é uma tarefa de aprendizado supervisionado.

O aprendizado semissupervisionado é um paradigma mais recente com o objetivo de combinar as forças de cada um dos outros dois tipos de aprendizagem. Nele parte dos dados estão classificados, mas a maior parte não está. O intuito é conseguir propagar as classes dos exemplos classificados para os que ainda não estão classificados.

Existem inúmeras técnicas de classificação existentes na literatura, algumas das mais usadas são: árvores de decisão, aprendizado baseado em instâncias, redes Bayesianas, redes neurais artificiais, máquina de vetores de suporte (CARNEIRO; ZHAO, 2017).

Formalmente no aprendizado supervisionado, dois conjuntos de dados são utilizados, sendo um de treino e outro de teste, sendo que os dados de treino já estão rotulados com a sua classe.

A base de treino pode ser formalmente descrita da seguinte maneira. Vamos considerar a base de treino com n elementos como $B_{treino} = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n) | 1 \leq y_i \leq C; 1 \leq i \leq n\}$. O par (x_i, y_i) representa o i-ésimo objeto do problema, sendo $x_i = \{a_1, a_2, ..., a_m\}$ um vetor que descreve os m atributos do objeto, e y_i sendo sua respectiva classe dentre as k possíveis. A base de teste por sua vez $B_{teste} = \{(x_{n+1}, y?), \ldots, (x_{n+j}, y?)$. O objetivo aqui é criar uma função $f : x \rightarrow y$ tal que possa determinar os rótulos dos objetos da base de teste, que, por sua vez, tem como objetivo testar a habilidade preditiva do classificador gerado (RESENDE; CARNEIRO, 2021).

2.2 Redes Complexas

O estudo de redes complexas tem atraído muita atenção nos últimos anos consistindo em uma ferramenta eficiente para modelagem de redes sociais, redes biológicas e de comunicação (COSTA et al., 2007). Redes complexas lidam com inúmeros tipos de situação, sendo comumente representadas como grafo. Um grafo é normalmente representado como $G(V, E)$ sendo V um conjunto de objetos que no caso são nossos vértices e E o conjunto

de arestas do grafo. Um exemplo de uma rede complexa é a WWW (*World Wide Web*) onde os nós são os documentos (páginas *web*) e as arestas são os *hiperlinks* (URL's) que apontam de um documento para outro (ALBERT; BARABáSI, 2002). A Figura 3 ilustra um exemplo de uma rede complexa em forma de grafo que demonstra a partir do ano, a relação dos artigos, autores, países e palavras-chave de uma base de dados.

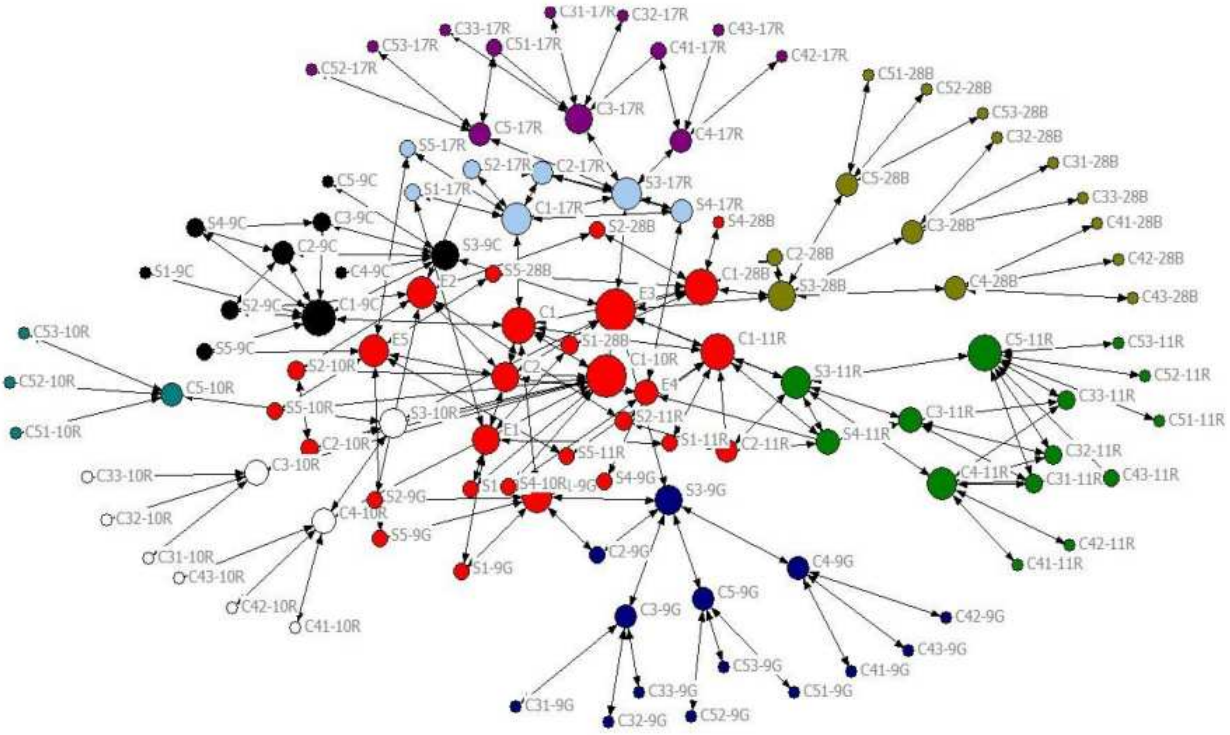


Figura 3 – Exemplo de uma rede complexa para demonstrar uma rede de comunicações da Operação Ágata. Retirada de Filho e Moura (2018).

Medidas de redes complexas são utilizadas para modelar cálculos em diferentes aplicações, oriundas de estudos do campo da estatística, sistemas complexos, matemática, sistemas não lineares, entre outros (COSTA et al., 2007). Elas facilitam o reconhecimento de certos tipos de propriedades ou grupos através da análise dos dados em rede.

2.2.1 Assortividade

É a medida que quantifica a tendência dos vértices se conectarem. O coeficiente de assortatividade pode assumir valores entre $[-1, 1]$, de modo que os positivos indicam que pares de vértices diretamente conectados tendem a se comportar da mesma forma, enquanto os negativos indicam uma maior probabilidade de que os vértices conectados tenham comportamento distintos (CARNEIRO; ZHAO, 2017). Seja $L = |E|$ o número de conexões na rede e i_u, k_u o grau dos vértices i e k que compõem uma aresta u, a

assortatividade da rede pode ser calculada por:

$$r = \frac{L^{-1}\sum_u i_u k_u - [L^{-1}\sum_u \frac{1}{2}(i_u + k_u)]^2}{L^{-1}\sum_u \frac{1}{2}(i_u^2 + k_u^2) - [L^{-1}\sum_u \frac{1}{2}(i_u + k_u)]^2}. \quad (1)$$

2.2.2 Coeficiente de Agrupamento

Mede o grau com que os nós da rede tendem a agrupar-se. O coeficiente de um nó mede o quão próximo ele está de formar um grafo completo (clique), a equação é dada por:

$$CC_i = \frac{|e_{us}|}{k_i(k_i - 1)}, \quad (2)$$

onde $CC_i \in [0, 1]$, $|e_{us}|$ denota o número de conexões compartilhadas pelos vizinhos diretos do nó i e k_i é o grau do nó i. Portanto, o coeficiente de agrupamento da rede pode ser obtido por:

$$CC = \frac{1}{N}\sum_{i=1}^{N} CC_i, \quad (3)$$

2.2.3 Modularidade

A modularidade faz o cálculo Q que varia entre 0 e 1, onde 0 representa uma rede totalmente aleatória e 1 no qual indica comunidade claramente dividas. É utilizada no contexto de detecção de comunidades. Formalmente é dada por:

$$Q = \frac{1}{2E}\sum_{u,v}[e_{uv} - \frac{k_u k_v}{2E}]\delta(c_u, c_v), \quad (4)$$

onde E é o número total de nós da rede, e_{uv} indica a fração de arestas que conectam vértices de u para v na comunidade, k_u denota o grau do vértice u e $\delta(c_u, c_v)$ representa o delta de *Kronecher*, o qual produz 1 se $c_u = c_v$ e 0, caso contrário. A modularidade é capaz de mensurar a qualidade da divisão feita na rede.

2.2.4 Proximidade

A proximidade mede o inverso do caminho mínimo médio entre um vértice e todos os outros da rede (CARNEIRO; ZHAO, 2017). A fórmula é dada por:

$$AC_i = \frac{n-1}{\sum_{j=1}^{n} d(i,j)}, \tag{5}$$

onde $d(i,j)$ é o caminho mínimo entre os vértices i e j, n é a quantidade de vértices da rede. A proximidade média é dada por:

$$AC = \frac{1}{n}\sum_{i=1}^{n} AC_i. \tag{6}$$

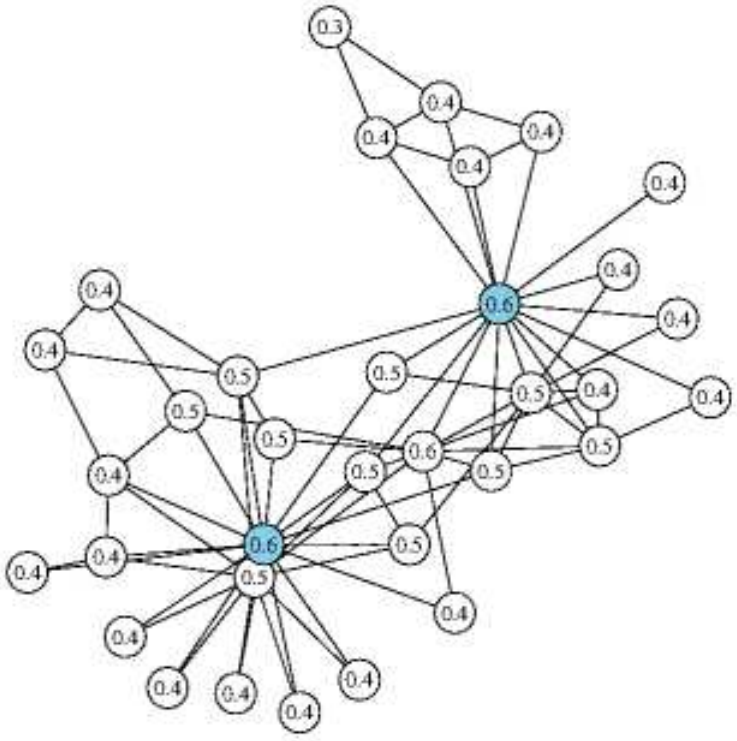


Figura 4 – Proximidade normalizada dos objetos da rede "Caratê" retirada de (CARNEIRO; ZHAO, 2017).

2.2.5 PageRank

É a medida que mede a importância de um nó contabilizando a quantidade de *links* (arestas) apontando para ele. Pode ser interpretada como um modelo de caminho aleatório, o qual assume que um agente está passeando sobre o grafo. Em cada nó, o agente,

aleatoriamente, seleciona uma das conexões de saída e vai para aquele nó vizinho. Além disso, o agente pode saltar, aleatoriamente, para qualquer outro nó do grafo, isso assegura que o processo de caminhada não ficará preso em um subgrafo sem conexões de saída (CARNEIRO; ZHAO, 2017). O *PageRank* pode ser descrito pela fórmula:

$$PR_j^{(t+1)} = \sum_{i \to j} \beta . \frac{PR_i^{(t)}}{k_i^{out}} + (1 - \beta)\frac{1}{n}, \tag{7}$$

onde $PR_j^{(t+1)}$ é o PageRank do vértice i, k_i^{out} é o grau de saída do vértice i, n é o número de nós na rede, β é a probabilidade de saltos aleatórios, $i \to j$ denota a conexão do nó i para o nó j.

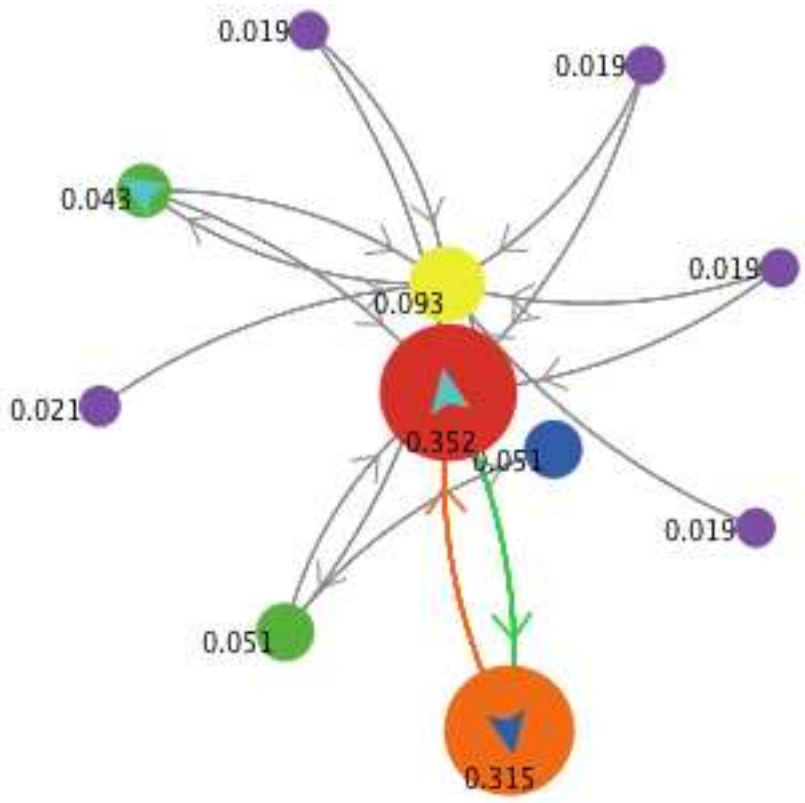


Figura 5 – Valores de PageRank para os vértices.
Fonte: https://ccl.northwestern.edu/netlogo/models/PageRank

2.3 Aprendizado de máquinas em redes complexas

Nesta seção, serão discutidas e apresentadas algumas técnicas para a construção da rede bem como para aprendizado supervisionado em redes.

2.3.1 Construção da Rede

Uma etapa crucial do aprendizado de máquina baseado em rede é a construção do grafo. Diferentemente das técnicas tradicionais a etapa de treinamento aqui envolve, principalmente, a formação da rede a partir das conexões entre os vértices, onde os nós e arestas representam, respectivamente, os objetos e as relações que existem entre cada nó conectado pela aresta (CARNEIRO; ZHAO, 2017). Nas próximas subseções serão discutidos alguns dos métodos para a criação de uma rede.

2.3.1.1 Rede completamente conectada

Neste método, é criado um grafo totalmente conectado, ou seja, há uma aresta para todo par de vértice. O grafo necessita que as arestas tenham pesos para que nós similares tenham uma aresta com peso alto entre eles. A vantagem de utilizar é o aprendizado a partir do peso, a desvantagem é o custo computacional para problemas muito complexos, como a rede construída é muito densa, isso pode ser um fator que seria impossível o uso em alguns casos. Na literatura há conceitos relacionados a formação de rede, alguns serão apresentados nas próximas subseções. Entretanto, alguns trabalhos apontam empiricamente que redes completamente conectadas apresentam pior desempenho que redes esparsas (ZHU; LAFFERTY; ROSENFELD, 2005).

2.3.1.2 Rede k vizinhos mais próximos

Nesta abordagem, a construção do grafo é feita a partir de relação de uma instância com outra instância, onde a forma de se construir a rede k vizinhos mais próximos (kNN) tem sido amplamente usada na literatura (OZAKI et al., 2011). Os nós i e j são conectados por uma aresta se i está na lista de vizinhos mais próximos de j ou vice versa. k é um parâmetro ajustável que controla o número de conexões dos vértices da rede, ou seja, quanto maior o valor de k, mais densa a rede se torna por conta do número de conexões. Um valor de k muito baixo pode acabar resultando em grafos desconexos. Os métodos baseados em k vizinhos, normalmente, produzem grafos esparsos o que se torna uma vantagem em relação aos grafos completamente conectados, porém nesse caso o desafio passa a ser determinar o melhor valor para k (CARNEIRO; ZHAO, 2017).

2.3.1.3 Rede vizinhança de raio ε

Esse método de vizinhança de raio ε (εN) constrói uma rede não direcionada, onde cada aresta consiste em um par de vértices (v_i, v_j) (CARNEIRO; ZHAO, 2017), que segue a seguinte fórmula:

$$e_{ij} = \begin{cases} 1, & S_{ij} >= \varepsilon \\ 0, & \text{caso contrário.} \end{cases}$$

Assim, o valor ε possibilita o controle do raio da vizinhança, tornando a identificação do valor ótimo para ε uma questão desafiadora.

2.3.2 Aprendizado supervisionado em redes complexas

A classificação baseada em redes complexas é um tópico bastante recente, mas que já obteve resultados interessantes, especialmente por revelar vantagens em certas ocasiões sobre os algoritmos tradicionais, tais como a ausência de parâmetros, a habilidade para detectar classes de diferentes formas, e a possibilidade de mapear relações espaciais, topológicas e funcionais dos dados (CARNEIRO; ZHAO, 2017).

Como o foco deste trabalho é o aprendizado supervisionado em redes complexas, nas próximas subseções serão discutidos trabalhos correlatos sobre classificação de dados utilizando redes complexas.

2.3.2.1 Grafo K-associados ótimo para classificação de dados

Em relação às outras técnicas, essa técnica se caracteriza por: (1) Uma nova medida denominada *pureza* que mede o nível de mistura entre os componentes de classes distintas; e (2) ausência de seleção de parâmetros (BERTINI et al., 2011).

Na fase de treinamento, ou construção da rede, os dados em forma de vetor de atributos são representados como um grafo k-associados, que é um grafo com k vizinhos mais próximos no qual as conexões existem apenas entre os vizinhos que possuem a mesma classe e que satisfazem as restrições de pureza. A Figura 6 apresenta o cálculo de pureza para três componentes (CARNEIRO; ZHAO, 2017).

Basicamente, a rede k-associados ótima é formada a partir de uma estimativa para o valor de k para cada componente, visando maximizar o nível de pureza dos componentes da rede, a partir desse princípio, para cada valor de k, é calculada a pureza e usada para comparar com o valor de pureza de outros grafos k-associados gerados, de tal forma que o componente com maior pureza é mantido. O grafo k-associados ótimo é construído primeiramente com o valor de k sendo 1, a partir dele se inicia um laço cuja condição de parada é uma heurística que visa garantir que o algoritmo só pare quando ele alcançar o grafo com grau máximo de pureza (BERTINI et al., 2011).

Na fase de classificação, já obtida a rede k-associados ótima, a classificação de novas instâncias é possível. A rede ótima é utilizada por um classificador ótimo de Bayes, de forma que a classificação acontece por meio de um cálculo de probabilidade do novo item pertencer a classe.

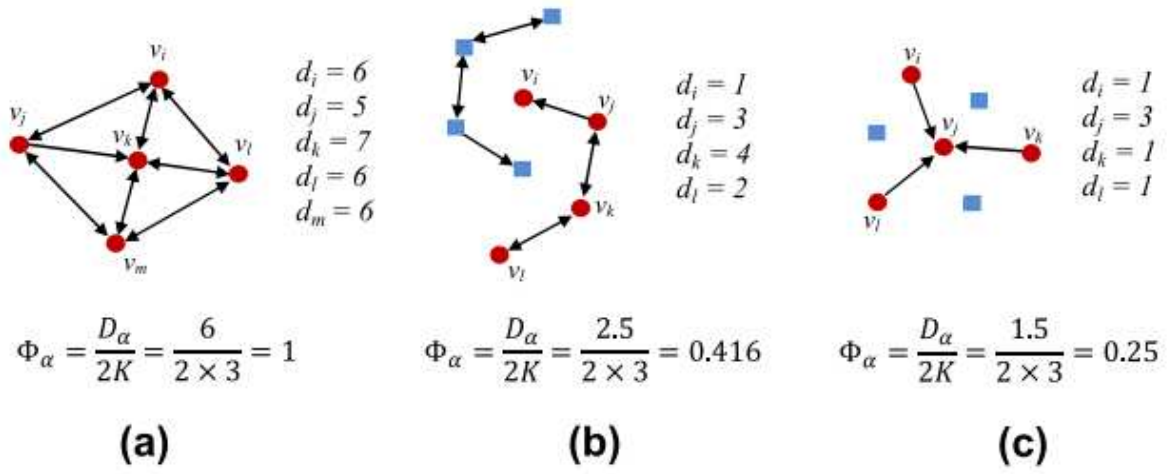


Figura 6 – Medida de pureza calculada com 3 níveis de mistura. Figura retirada de (BERTINI et al., 2011).

2.3.2.2 Classificação via conformidade de padrão

A classificação via conformidade de padrão é uma técnica de classificação de alto nível que pode ser dividida em dois passos: treinamento e classificação. Na etapa de treinamento os objetos classificados são utilizados para a criação da rede. Na classificação, a técnica utiliza de medidas de redes complexas para realizar cálculos antes e após a inserção dos itens de teste. Caso haja grande variação no cálculo dessas medidas para uma dada classe, diz-se que o item de teste não tem conformidade com o padrão da classe. Caso seja pequena a variação, possivelmente há conformidade de padrão entre a nova instância e os dados da classe (CARNEIRO; ZHAO, 2017).

A técnica de alto nível proposta em Silva e Zhao (2012) consiste em um *framework* que combina classificadores de baixo e alto nível. O que difere essa técnica das convencionais nos quais são chamadas de classificadores de baixo nível, é que a classificação de alto nível não considera apenas os atributos físicos dos dados, mas também outras características como padrão de formação, no qual são utilizadas medidas de redes complexas para se obter. Portanto, essas medidas são empregadas a fim de aumentar a acurácia dos métodos tradicionais (CARNEIRO; ZHAO, 2017).

De acordo com o critério, um vértice i será conectado utilizando εN quando o número de conexões é maior que o valor de k, caso seja menor ou igual o valor de k será utilizado o kNN. Com a rede construída com base nos critérios apresentados, poderemos utilizá-la para realizar a classificação de novas instâncias.

Na classificação de alto nível, é examinado o impacto que a inserção de um objeto de teste y tal que $y \in X_{test}$ para cada classe isoladamente. A variação dos resultados das

medidas de rede irão indicar a qual classe y é pertencente. Dado um grafo G_c se a inserção de y causar pequena variação nas medidas de rede, a classificação retorna um valor alto de associação de y para a classe c, caso contrário, retorna um pequeno valor de associação para a classe c, indicando que o item de teste, provavelmente, não pertence àquela classe.

A classificação de um item de teste y de acordo com a técnica proposta em Silva e Zhao (2012) pode ser descrita como:

$$\mathcal{M}_y^{(c)} = (1-\lambda)\mathcal{C}_y^{(c)} + \lambda\mathcal{H}_y^{(c)}, \tag{8}$$

onde $\mathcal{M}_y^{(c)}$ representa a associação produzida pelos algoritmos de baixo e alto nivel quando y é avaliado para a classe c. λ é o parâmetro de combinação entre os dois classificadores.

A classificação de alto nível entretanto, classifica um item de teste y para uma classe c de forma:

$$\mathcal{H}_y^{(c)} = \frac{\sum_{u=1}^{Z} \delta(u)[1 - f_y^{(c)}(u)]}{\sum_{g \in L} \sum_{u=1}^{Z} \delta(u)[1 - f_y^{(g)}(u)]}, \tag{9}$$

onde $\mathcal{H}_y^{(c)}$ representa as medidas de redes empregadas pelo algoritmo de alto nível, as outras variáveis são parâmetros para indicar a influência das medidas de rede no processo de classificação. A função $f_y^{(g)}$ retorna uma resposta se o item de teste y apresenta ou não conformidade com os padrões da classe c.

Em Carneiro et al. (2023), foi feito um estudo utilizando exames de eletroencefalograma que é um exame que analisa a atividade elétrica cerebral espontânea, captada através da utilização de eletrodos colocados sobre o couro cabeludo, de pessoas em coma tratadas na unidade de tratamento intensivo do Hospital de Clinicas da UFU, com esses exames foram extraídas as características dos sinais do eletroencefalograma, construído uma rede utilizando redes complexas utilizando das mais avançadas e modernas medidas de redes para classificação de alto nível, combinando o classificador de alto nível com outras técnicas de aprendizagem supervisionadas conhecidas. O resultado obtido deste estudo foi a relevância da informação estrutural e topológica capturado pelas medidas de redes na análise do eletroencefalograma, foi relatado também a importância de se extrair as características relevantes para a classificação dos dados.

2.3.3 Classificação via caracterização de importância

Em Carneiro e Zhao (2018), é apresentado um método de classificação de alto nível, onde o algoritmo realiza a classificação considerando a importância de cada objeto da rede, fazendo com que as medidas de redes complexas classifique um novo objeto aonde recebeu

o maior valor de importância. Para o cálculo do valor da importância foi utilizado como base o algoritmo PageRank, método utilizado pelo Google para classificar a importância das páginas para o resultado da busca.

Após a formação de uma rede $\mathcal{G}$ a partir de um conjunto de dados rotulados $\mathcal{X}_{train}$, suponha que uma instância de teste $y \in \mathcal{X}_{test}$, precisa ser classificada. Baseado em (CARNEIRO; ZHAO, 2017) a importância de um item de teste y, detonada por $\mathcal{I}$, em relação à classe $l \in \mathcal{L}$ é dada por:

$$\mathcal{L}_y^{(l)} = \sum_{j \in \Lambda_y^{(l)}} \mathcal{I}_j, \tag{10}$$

onde $j \in \mathcal{X}_{train}$ denota um vértice rotulado, $\Lambda_y^{(l)}$ é o conjunto de nós que pertencem à classe l na qual y é temporariamente conectado, e $\mathcal{I}_j$ representa a importância do vértice j. De maneira geral, a Eq. 10 calcula o valor de importância da instância de teste y em relação a cada classe l. Em sequência, a classe atribuída à y é aquela em que ele recebe o maior valor de importância, i.e.

$$\varphi = \underset{l \in \mathcal{L}}{argmax}\ \mathcal{I}_y^{(l)}. \tag{11}$$

A importância de um nó $j \in \mathcal{X}_{train}$ é quantificada por iterar o seguinte sistema, o qual é equivalente à formulação do PageRank:

$$\mathcal{I}_j^{(t+1)} = \sum_{i \to j} \beta \cdot \frac{\mathcal{I}_i^{(t)}}{k_i^{out}} + (1-\beta)\frac{1}{n}, \tag{12}$$

onde $i \to j$ representa uma aresta do nó i para o nó j, $\mathcal{I}_i$, refere-se à importância do nó i, k_i^{out} represetan o grau de saída do nó i, e n é o número total de nós da rede. β significa a probabilidade de saltos aleatórios, a qual é definida como 0.85 (BOLDI; SANTINI; VIGNA, 2005).

A seguir é definido o conjunto de conexões temporárias para y, detonado por $\Lambda_y^{(l)}$ na Eq. 10, o qual é obtido a partir da seguinte formulação:

$$\Lambda_y^{(l)} \cup \{j \mid \mathcal{F}_{y,j} \geq 0\ e\ j\ \in\ l\}, \tag{13}$$

onde $\mathcal{F}_{y,j}$ representa a eficiência diferencial espaço-estrutural é uma função que avalia se uma conexão entre um nó y e outro nó j contribui para o aumento ou diminuição da

eficiência do componente que inclui o nó j. Caso a conexão eleve a eficiência, o nó j é incorporado ao conjunto $\Lambda_y^{(l)}$. Por outro lado, se nenhuma conexão para j resulta em um aumento da eficiência em qualquer ligação para y, i.e., $\mathcal{F}_{y,j} < 0$ para todo j, são incluídos no conjunto $\Lambda_y^{(l)}$ aqueles nós cujas ligações com y têm o menor impacto negativo na eficiência do componente.

A eficiência calculada é derivada do menor caminho entre todos os pares de vértices em um componente e será definida a seguir como demonstrado em Carneiro e Zhao (2017). A formulação matemática de $\mathcal{F}_{y,j}$ é dada por:

$$\mathcal{F}_{y,j} = \mathcal{E}_{j\in\alpha}^{(\alpha)} \cdot \gamma - D_{y,j}, \tag{14}$$

onde $\mathcal{E}^{(\alpha)}$ é a medida de eficiência do componente α, a qual é dada por:

$$\mathcal{E}^{(\alpha)} = \frac{1}{n^{(\alpha)}} \sum_{i\in\alpha} \xi_i^{(\alpha)}, \tag{15}$$

e $D_{y,j}$ refere-se à distância Euclidiana entre itens de dados y e j na forma de vetor. Enquanto o primeiro termo tem o objetivo de capturar propriedades estruturais dos dados, o segundo termo captura as características físicas dos dados. Portanto, a medida de eficiência diferencial espaço-estrutural combina características físicas e topológicas embutidas nos dados de entrada.

2.3.4 Trabalhos relacionados

Há trabalhos recentes que realizam a análise das medidas de redes e combinação de algoritmos, fazendo isso para a classificação via conformidade de padrão. São eles:

- Em Resende e Carneiro (2021) os algoritmos e as medidas são investigados para a classificação multirrótulo para desenvolver uma técnica que combine algoritmos de classificação multirrótulo tradicionais com medidas de redes complexas.
- Em Carneiro, Gama e Ribeiro (2021) é feito um estudo que faz a medição do desempenho preditivo de algumas medidas de redes complexas.

Apesar de ambos os estudos apresentados realizarem análises de caracterização de medidas de redes complexas para classificação de dados, é interessante observar que eles estão direcionaods para a classificação via conformidade de padrão. Logo, esse tipo de análise para a caracterização de importância é uma lacuna importante da literatura, a qual pretende-se abordar com esse estudo.

Capítulo 3

Materiais e Métodos

Neste capítulo será discutido cada uma das bases de dados que utilizamos nas simulações sendo reais ou artificiais, assim como a técnica de classificação de alto nível via importância e as medidas de redes investigadas. Assim, a Seção 3.1 apresenta quais materiais foram selecionadas para o ambiente definido para se efetuar os experimentos. A Seção 3.2 discute a técnica de classificação de alto nível via importância e por fim a Seção 3.4 apresenta o ambiente experimental utilizado para a realização dos testes.

3.1 Materiais

Para os testes de classificação de dados foram utilizadas 2 tipos de bases: artificiais e reais. As bases artificiais foram geradas considerando 3 níveis de ruídos tornando a análise dos resultados a partir das medidas mais reais.

3.1.1 Bases de dados artificiais

Foram utilizadas as bases artificiais a partir de 2 domínios listados a seguir o *Two Moons* e o *Classification*, ambos com 3 diferentes níveis de ruídos sendo baixo, médio e alto.

- **Domínio *Classification***
  Este domínio representa conjuntos de dados com classes ou grupos de objetos seguindo distribuição Gaussiana. A Figura 7 demonstra graficamente as bases geradas com seus diferentes níveis de ruídos.

- **Domínio Two moons**
  Este domínio tem como característica 2 semicírculos intercalados que representa duas classes distintas. É um conjunto simples de visualização utilizando algoritmos de agrupamento e classificação, a Figura 8 demonstra graficamente as bases geradas com seus diferentes níveis de ruídos.

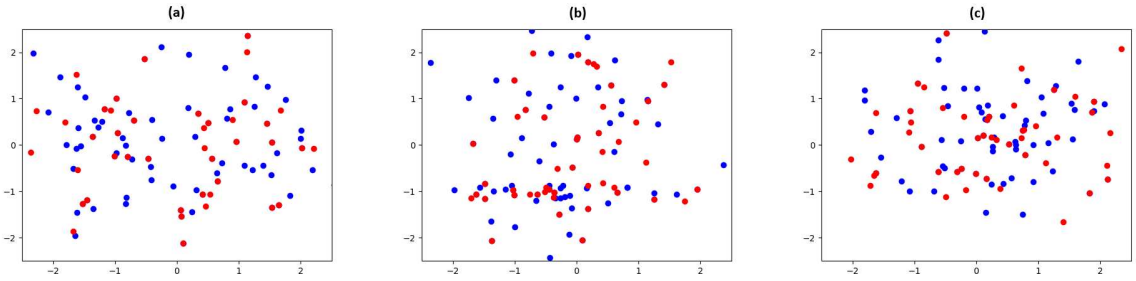


Figura 7 – Visualização dos dados da base artificial *classification* com 3 níveis de ruído. (a) nível baixo. (b) nível médio. (c) nível alto.

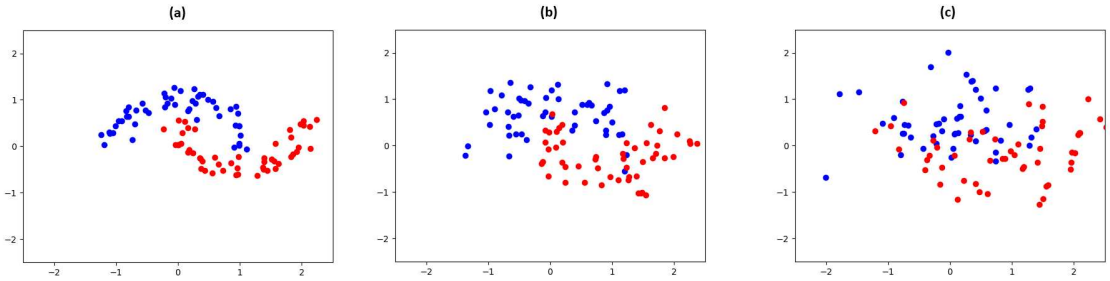


Figura 8 – Visualização dos dados da base artificial *two moons* com três níveis de ruído. (a) nível baixo. (b) nível médio. (c) nível alto.

3.1.2 Bases de dados reais

Foram selecionadas 4 bases de dados reais para os testes sendo eles: Appendicitis, Iris, Glass identification e Sonar. Sendo as quatro disponíveis no UCI Machine Learning Repository. A Tabela 1 demonstra as características de cada base.

Tabela 1 – Descrição das propriedades das bases reais utilizadas no trabalho (*Appendicitis*, *Glass identification*, *Iris* e *Sonar*), para cada base tem-se o número de instâncias, a quantidade de atributos e número de classes.

Nome	Instâncias	Atributos	Classes
Appendicitis	106	7	2
Glass	214	9	7
Iris	150	4	3
Sonar	208	60	2

- **Appendicitis**

  Neste conjunto de dados, os atributos representam sete medidas médicas tomadas em 106 pacientes, onde a classificação diz se o paciente tem apendicite em binário, sendo 1 para positivo e 0 para caso negativo.

- **Glass Identification**

  Na base de dados de identificação de vidros temos 214 instâncias cada uma tendo

9 atributos sendo eles o índice de refração, quantidade de sódio, quantidade de magnésio, quantidade de alumínio, quantidade de silicone, quantidade de potássio, quantidade de cálcio, quantidade de bário e quantidade de ferro. Com base nos 9 atributos cada instância pode ser identificada em 7 tipos de vidro sendo vidro de construção flotado processado, vidro de construção não flotado processado, vidro de veículo flotado processado, vidro de veículo não flotado processado, recipientes, louças e faróis. O uso dessa base para estudos foi motivada pela investigação criminal, onde na cena do crime o vidro pode ser deixado como evidência se for corretamente identificado (GERMAN, 1987).

- **Iris**

  Nesta base de dados temos que cada instância representa uma planta, onde cada planta contém 4 atributos o comprimento da sépala em centímetros, a largura da sépala em centímetros, o comprimento da pétala em centímetros e a largura da pétala também em centímetros, podendo cada planta ser classificada em uma das 3 classes *Íris Setosa*, *Íris Versicolour* e *Íris Virginica*. É uma das bases de dados pioneiras na literatura em termos de métodos de classificação, amplamente utilizada em estatísicas e aprendizado de máquina (FISHER, 1988).

- **Sonar**

  Essa base de dados contém padrões obtidos ao refletir sinais de sonar em um cilindro de metal em vários ângulos e sob várias condições. O sinal do sonar transmitido é um sinal sonoro modulado em frequência, aumentando a frequência. Cada padrão é um conjunto de 60 números que representam a energia dentro de uma banda de frequência específica. A classificação se dá pelas letras R se o objeto for uma rocha e M se for uma mina (cilindro de metal) (SEJNOWSKI; GORMAN, 1988).

3.2 Técnica de classificação via importância e medidas de rede

3.2.1 Classificação de dados via caracterização de importância

A classificação de dados via caracterização de importância tem como objetivo, considerar não apenas as características físicas, mas também considerar o nível de importância que um nó tem dentro da rede. Como dito anteriormente as técnicas tradicionais assumem que todos os objetos têm a mesma relevância que os outros. O algoritmo usado para calcular a importância é derivado do *PageRank*, utilizado pelo sistema de busca da Google.

Suponha que $X_{train} = x_i, i = 1, ..., n$ um conjunto de dados onde cada item x_i tenha apenas uma classe associada, e $X_{test} = x_i, i = 1, ..., m$ um conjunto de dados onde pre-

cisamos rotular. Cada item tem um conjunto d de atributos tal que $x_i = x_{i1}, ..., x_{id}$ e $X_{train} \cap X_{test} = \emptyset$, podemos dividir a técnica em duas etapas, treinamento e classificação (CARNEIRO; ZHAO, 2018).

Treinamento:

- ❏ A técnica proposta utiliza um método de formação de rede como os discutidos na Seção 2.3 utilizando os dados a partir do conjunto X_{train}.
- ❏ Chamamos a rede formada de $G = V, E$, onde V é o conjunto de vértices e E o conjunto de arestas.
- ❏ Para a tarefa de classificação é necessário o uso do *PageRank* e da medida de $eficiência\ diferencial\ espaço - estrutural$ descrita na Subseção 2.3.3, as quais são obtidas a partir de G (CARNEIRO; ZHAO, 2017).

Classificação:

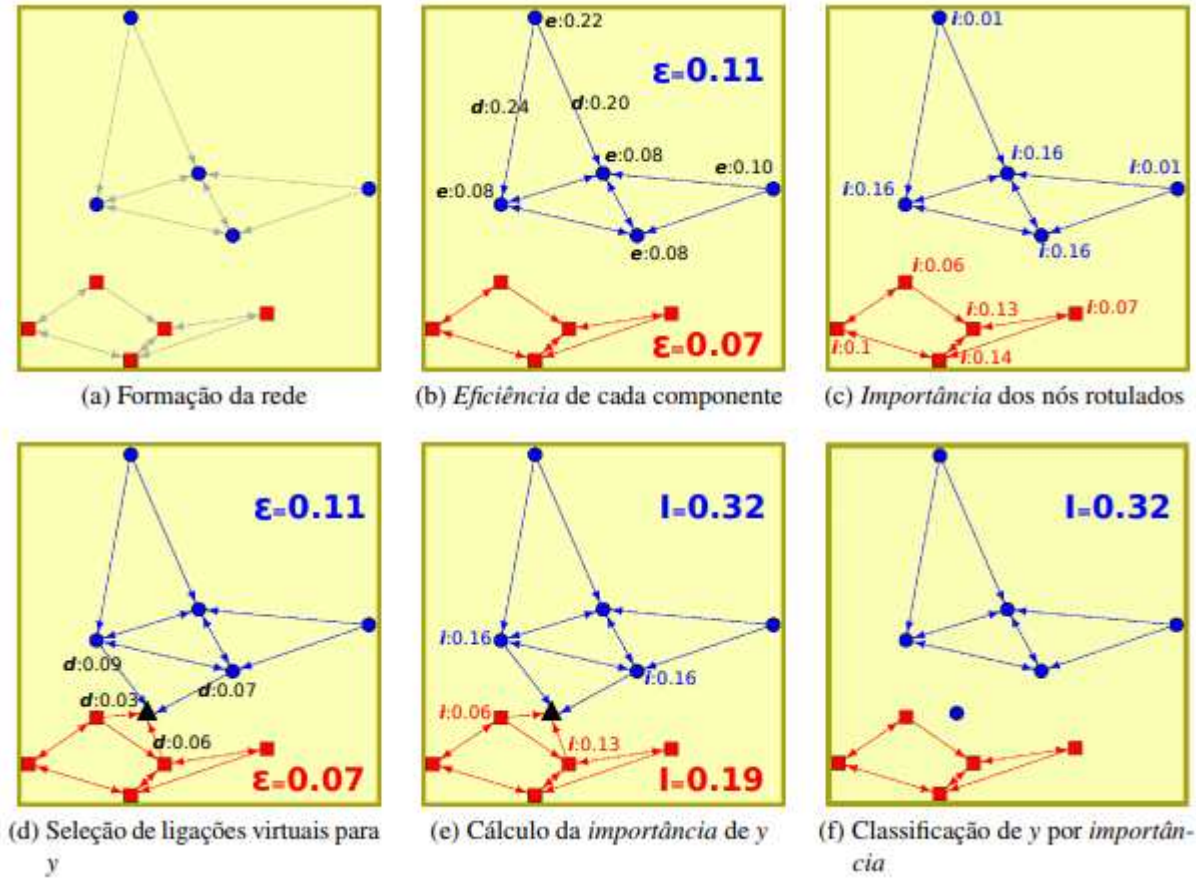
- ❏ Um novo item sem rótulo $y \in X_{test}$ é apresentado ao classificador.
- ❏ O item é conectado a um conjunto de vértices baseado na medida de $eficiência\ diferencial\ espaço - estrutural$ dos componentes presentes na rede.
- ❏ Para predizer a classe do objeto y é calculado a importância dele em cada classe a partir das conexões temporárias estabelecidas.
- ❏ Por fim, y é atribuído a classe no qual recebeu um maior nível de importância.

A Figura 9 ilustra passoa a passo o processo de treinamento e classificação da técnica. A formação de rede é mostrada na Figura 9a; o cálculo da eficiência de cada nó é ilustrada na Fig. 9b; o valor da importância de cada nó rotulado é mostrado na Figura 9c; Após isso, quando um objeto y de teste é conectado a um conjunto de nós temporariamente de acordo com a medida de $eficiência\ diferencial\ espaço - estrutural$ Figura 9d; Logo após, o valor de importância que y recebe a partir das classes dos nós conectados é exibido na Figura 9e; a classificação do objeto y na classe onde apresentou o maior valor de importância é descrito pela Figura 9f.

3.2.2 Medidas de redes complexas

Das medidas de redes utilizadas para as análises dos resultados utilizamos 4 medidas sendo:

- ❏ Proximidade.
- ❏ Grau.



(a) Formação da rede (b) *Eficiência* de cada componente (c) *Importância* dos nós rotulados

(d) Seleção de ligações virtuais para *y* (e) Cálculo da *importância* de *y* (f) Classificação de *y* por *importância*

Figura 9 – Demonstração passo a passo do treinamento e classificação da técnica, retirado de (CARNEIRO; ZHAO, 2017)

- *PageRank*.
- Menor caminho médio.

3.3 Medidas de avaliação

- **Acurácia**

  Faz a contagem dos dados registrados corretamente ou incorretamente nas classes, resultando assim na porcentagem de acertos dos dados de teste que foram classificados corretamente. Seja um conjunto X de dados de teste, C o número de objetos classificados corretamente, temos:

$$A_c = \frac{C}{|X|}100 \tag{16}$$

❑ **Desvio padrão**

Indica o grau de dispersão do conjunto de dados, portanto, indica a semelhança do conjunto. Quanto mais distante de 0 for o desvio padrão, mais diferentes são os dados. Dado X_y o valor do objeto y, M_a a média aritmética dos dados e n o número de dados, temos:

$$D_p = \sqrt{\frac{\sum_{i=1}^{n} (X_i - M_a)^2}{n}} \tag{17}$$

3.4 Ambiente experimental

A implementação do algoritmo para reproduzir a técnica de classificação baseada em importância proposta em Carneiro e Zhao (2017) foi feita na linguagem Python assim como suas bibliotecas e pacotes existentes. Para o treinamento do algoritmo foi utilizada validação cruzada estratificada em 10 vezes. Nessa validação, o conjunto de dados é dividido em 10 pastas, onde em cada execução, 9 pastas são utilizadas para treino e 1 deixada para o teste. Além disso, o número de repetições utilizado foram 3, totalizando assim 30 execuções.

Os parâmetros considerados nas análises foram um valor γ variando entre 0, 0.5, 1 e 2, e o valor de K no que diz respeito à construção da rede KNN (K-vizinhos mais próximos) para avaliar a influência da rede no desempenho das medidas. O desempenho das medidas é avaliado em termos de acurácia média e desvio padrão sobre as bases artificiais e reais apresentadas neste capítulo.

3.4.1 Equipamento utilizado

Para os testes foi o utilizado um computador nas seguintes configurações.

❑ Processador Ryzen 7 5800X

❑ 32 GB de memória ram a 3600MHZ

❑ Placa de vídeo RTX 3060

3.4.2 Bibliotecas utilizadas

Para o desenvolvimento do algoritmo foram utilizadas as bibliotecas do Python a seguir.

❑ Numpy

- ❏ Igraph
- ❏ Sklearn
- ❏ Matplotlib

Capítulo 4

Resultados Experimentais

Neste capítulo, os resultados obtidos serão apresentados, acompanhados de análises sobre o desempenho de cada medida de rede. Serão consideradas tanto as bases artificiais quanto as reais anteriormente delineadas, além de explorar o comportamento variável das medidas de redes complexas em função da acurácia

4.1 Resultados em bases artificiais

Domínio Classification:

A Tabela 2 apresenta o conjunto de resultados com a melhor acurácia de acordo com a análise nos testes variando tanto o valor K na construção da rede quanto o valor de γ. Podemos observar que em bases com níveis de ruídos menores o valor de γ menor resultou em melhores resultados, enquanto que em bases com níveis de ruídos maiores o aumento no valor de γ contribuiu em uma melhoria na acurácia.

Tabela 2 – Melhores valores de acurácia e seu respectivo desvio padrão das medidas de redes utilizadas para o domínio *Classification*, considerando todos os valores de γ (0, 0.5, 1, 2) e K (1 ao 15).

	Classification (melhor)		
Medidas de redes	**Baixo ruído**	**Médio Ruído**	**Alto ruído**
Grau	**95±6.2** (k=8, γ=0)	87.6±9.2(k = 15,γ=0.5)	**78.3±13.2** (k=7, γ=1)
Proximidade	94.3±6.6(k = 1,γ=0)	86.3±10.8(k = 1,γ=2)	76.3±14.7(k = 4,γ=1)
PageRank	94.67±6.7(k = 2,γ=0.5)	88±9.4(k = 6,γ=1)	78±12.2(k = 9,γ=1)
Menor Caminho	94.6±6.7(k = 2,γ=0.5)	**89.3±9.1** (k=14, γ=0.5)	76±15.8(k = 9,γ=0.5)

A Figura 10 mostra que a variação de desempenho é pouco alterada pelo valor de K quando $\gamma = 0$. A Figura 11 demonstra que a variação do parâmetro de construção da rede K em valores maiores pode não ter uma influência impactante na variação da acurácia das medidas de redes ao considerar dados com baixo nível de ruído. Por outro lado, observa-se que para algumas medidas no nível médio de ruído pode haver um aumento nos resultados de acurácia, no entanto, considerarmos o nível alto de ruído nota-se que

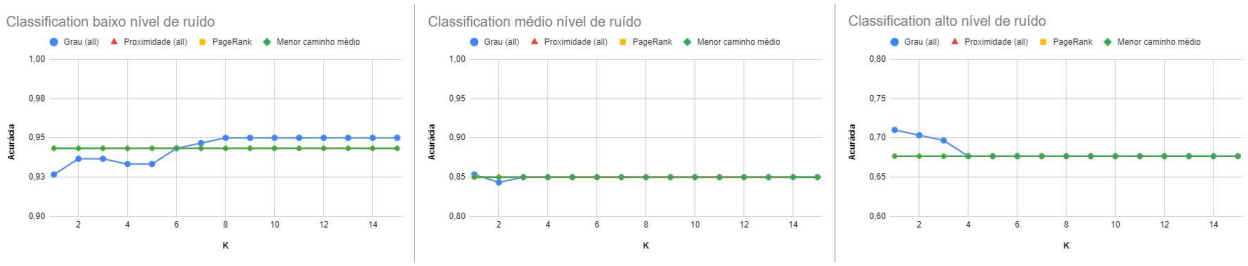


Figura 10 – Variação da acurácia de acordo com o aumento do valor de K com o valor de γ em 0 para as medidas de redes complexas no domínio *Classification*, cada uma indicando os resultados para os diferentes níveis de ruídos (baixo, médio e alto).

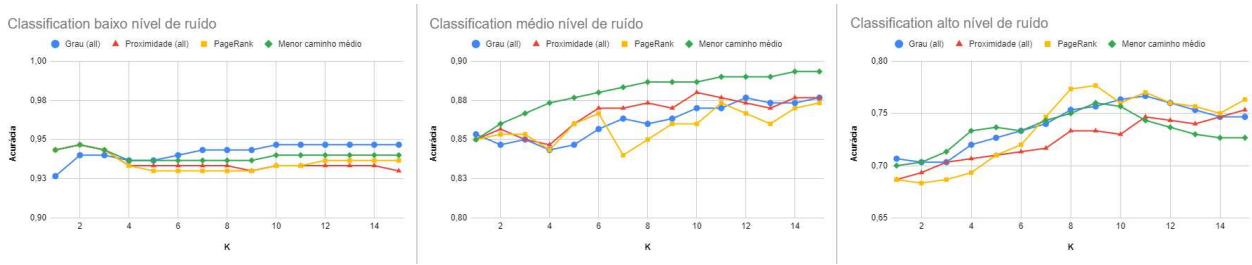


Figura 11 – Variação da acurácia de acordo com o aumento do valor de K com o valor de γ em 0.5 para as medidas de redes complexas no domínio *Classification*, cada uma indicando os resultados para os diferentes níveis de ruídos (baixo, médio e alto).

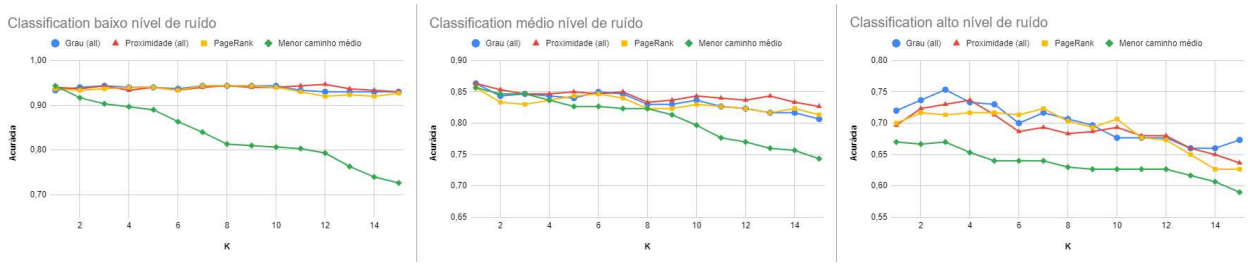


Figura 12 – Variação da acurácia de acordo com o aumento do valor de K com o valor de γ em 1 para as medidas de redes complexas no domínio *Classification*, cada uma indicando os resultados para os diferentes níveis de ruídos (baixo, médio e alto).

com o incremento de γ, valores muito altos de K podem causar uma queda nos resultados, tal como exibido pelas Figura 12 e Figura 13.

De modo geral, é possível ver que as medidas usadas para os testes obtiveram resultados mais semelhantes ao considerar valores baixos de γ. Ao observar para os dados na base com alto nível de ruído vemos que os resultados têm uma variação maior enquanto o valor de K aumenta. A Figura 13 mostra como a medida de menor caminho médio

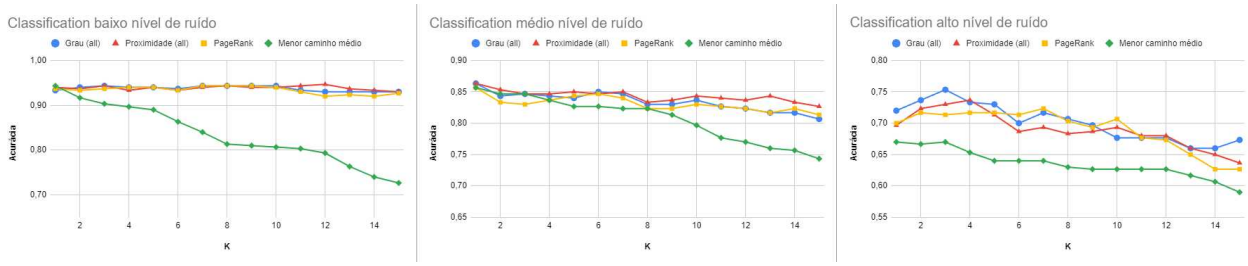


Figura 13 – Variação da acurácia de acordo com o aumento do valor de K com o valor de γ em 2 para as medidas de redes complexas no domínio *Classification*, cada uma indicando os resultados para os diferentes níveis de ruídos (baixo, médio e alto).

se torna mais sensível à variação de γ quando olhamos os dados e vemos a tendência de queda de acurácia enquanto o valor de K aumenta. Por outro lado as outras medidas de redes continuam com comportamentos semelhantes entre elas.

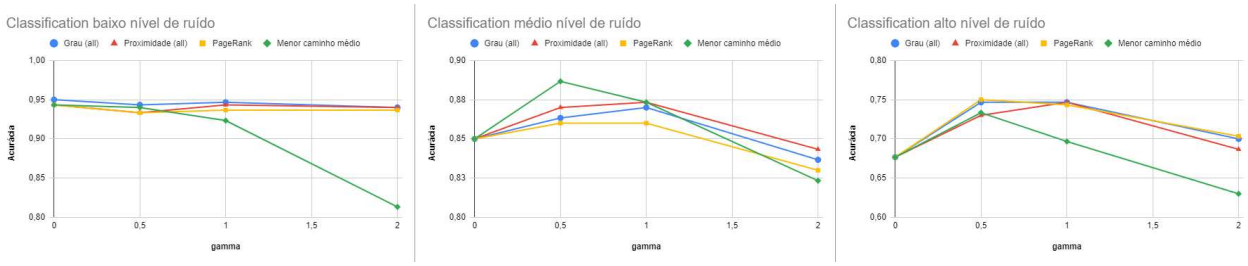


Figura 14 – Variação de acurácia de acordo com o aumento do valor de $\gamma \in \{0, 0.5, 1, 2\}$ para a base de dados *Classification* nos 3 níveis de ruído baseado nas médias de resultados da variação do valor de K do 1 ao 15.

A Figura 14 demonstra a sensibilidade de algumas medidas de redes para a variação do valor de γ, demonstrando que para baixo nível de ruído, os valores menores de γ passam a ser mais relevantes, especialmente para a medida menor caminho médio. Por outro lado, com médio e alto nível de ruído, é possível observar que o ajuste de valores intermediários de γ pode permitir ganhos consideráveis de desempenho preditivo.

Domínio Moons:

A Tabela 3 demonstra os resultados para as medidas de redes em termos de acurácia e desvio padrão para o domínio *two moons* variando o valor de K para a construção de rede e o valor γ. Observa-se que os resultados obtidos no nível de baixo ruído são bem similares como podemos ver na Figura 15. No entanto, observa-se que para níveis mais elevados de ruído, a acurácia diminui em face de valores menores de γ. Na Figura 16, com $\gamma = 0.5$, esse comportamento é parecido no cenário com baixo nível de ruído, embora alguma melhoria passe a ser perceptível para os níveis médio e alto. Por outro lado, quando aumentamos

o valor de γ para 1, obtemos valores de acurácia melhores como podemos observar na Figura 17, porém também temos uma queda no valor da acurácia nos níveis mais baixos quando o valor de K aumenta, o que acontece de modo ainda mais acentuado quando γ = 2, tal como exibido pela Figura 18. Também é possível observar que a medida menor caminho médio tem um desempenho preditivo menor em comparação às outras medidas quando o valor de γ aumenta enquanto as outras tendem a manter uma acurácia maior.

Tabela 3 – Melhores valores de acurácia e seu respectivo desvio padrão das medidas de redes utilizadas para o domínio *Two moons*, considerando todos os valores de γ (0, 0.5, 1, 2) e K (1 ao 15).

	Moons (melhor)		
Medidas de redes	**Baixo ruído**	**Médio Ruído**	**Alto ruído**
Grau	**99±3** (k=2, γ=0)	**95.3±7.1** (k=12, γ=0.5)	**79.3±12** (k=12, γ=1)
Proximidade	**99±3** (k=1, γ=0)	**95.3±8** (k=8, γ=1)	79±12(k = 14,γ=1)
PageRank	**99±3** (k=1, γ=0)	**95.3±10.3** (k=5, γ=1)	79±11.9(k = 3,γ=2)
Menor Caminho	**99±3** (k=1, γ=0)	94±7.5(k = 5,γ=0.5)	74±12(k = 14,γ=0.5)

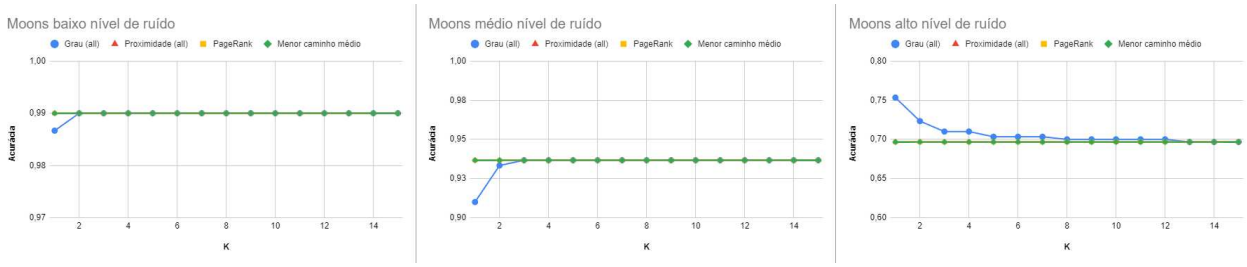


Figura 15 – Variação da acurácia de acordo com o aumento do valor de K com o valor de γ em 0 para as medidas de redes complexas no domínio *Two moons*, cada uma indicando os resultados para os diferentes níveis de ruídos (baixo, médio e alto).

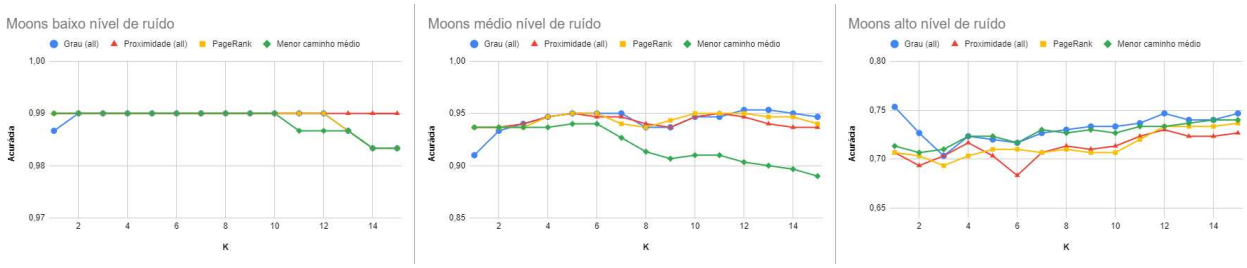


Figura 16 – Variação da acurácia de acordo com o aumento do valor de K com o valor de γ em 0.5 para as medidas de redes complexas no domínio *Two moons*, cada uma indicando os resultados para os diferentes níveis de ruídos (baixo, médio e alto).

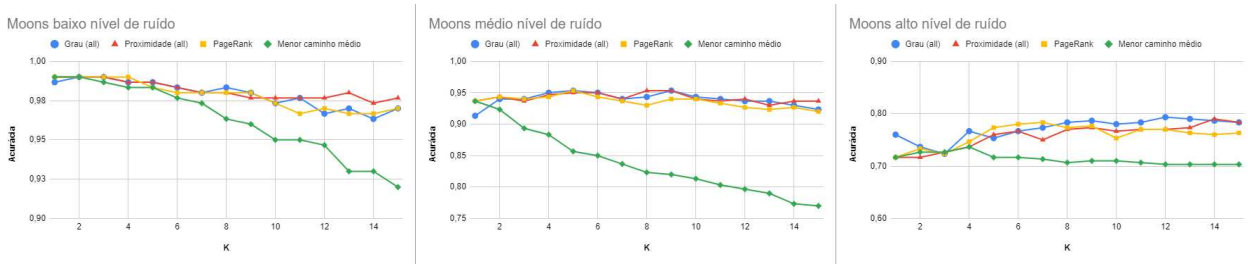


Figura 17 – Variação da acurácia de acordo com o aumento do valor de K com o valor de γ em 1 para as medidas de redes complexas no domínio *Two moons*, cada uma indicando os resultados para os diferentes níveis de ruídos (baixo, médio e alto).

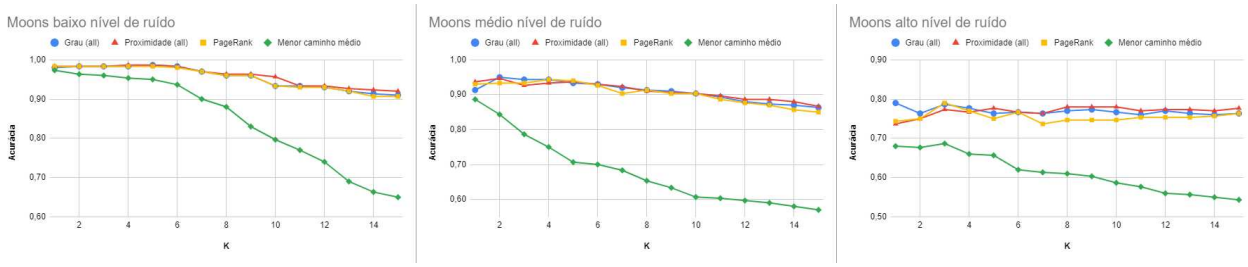


Figura 18 – Variação da acurácia de acordo com o aumento do valor de K com o valor de γ em 2 para as medidas de redes complexas no domínio *Two moons*, cada uma indicando os resultados para os diferentes níveis de ruídos (baixo, médio e alto).

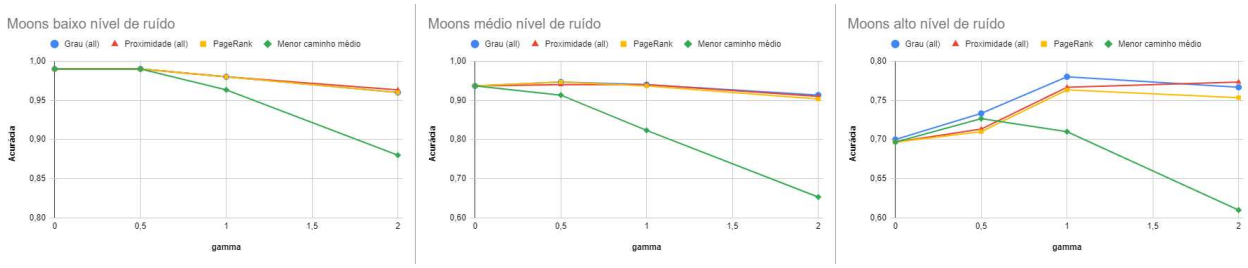


Figura 19 – Variação de acurácia de acordo com o aumento do valor de gamma $\gamma \in \{0, 0.5, 1, 2\}$ para a base de dados Moons nos 3 níveis de ruído baseado nas médias de resultados da variação do valor de K do 1 ao 15.

A Figura 19 demonstra novamente como foi na base de dados *Classification* que a medida de rede menor caminho médio demonstra ser mais sensível à variação do valor de γ.

Tabela 4 – Melhores valores de acurácia e seu respectivo desvio padrão das medidas de redes utilizadas para as bases de dados reais *Appendicitis*, *Glass identification*, *Iris* e *Sonar*. considerando todos os valores de γ (0, 0.5, 1, 2) e K (1 ao 15).

Medidas de redes	Appendicitis	Glass	Iris	Sonar
Grau	83±8.3(k = 8,γ=1)	72.6±8.6(k = 4,γ=0)	**98.4±2.8** (k=5, γ=1)	**84±7.0** (k=11, γ=0)
Proximidade	83±8(k = 1,γ=2)	73.8±7.7(k = 5,γ=2)	**98.4±2.8** (k=5, γ=1)	**84±7.0** (k=1, γ=0)
PageRank	**84±8.0** (k=13, γ=1)	73.2±8.7(k = 1,γ=0)	**98.4±2.8** (k=5, γ=1)	**84±7.0** (k=1, γ=0)
Menor Caminho	80±11(k = 1,γ=0)	**74.5±8.2** (k=9, γ=0.5)	96.2±5.3(k = 1,γ=0)	**84±7.0** (k=1, γ=0)

4.2 Resultados em bases reais

A Tabela 4 demonstra os resultados obtidos pelas medidas de redes complexas para as bases reais *Appendicitis*, *Glass identification*, *Iris* e *Sonar*. Esses resultados são os melhores encontrados de acordo com a variação dos parâmetros K e γ.

É possível observar baseado na Tabela 4 que todas as medidas utilizadas para os testes obtiveram resultados semelhantes quando olhamos base por base. Na base Appendicitis foi obtido o melhor resultado com 84% para a medida de rede PageRank. Na base Glass foi de 74.5% com a medida de menor caminho. Na base Iris temos um resultado de 98.4% de acurácia com as medidas de grau de entrada e saída, proximidade de entrada e saída e PageRank. E por fim a base Sonar obtivemos 84% de acurácia em todas as medidas de redes analisadas.

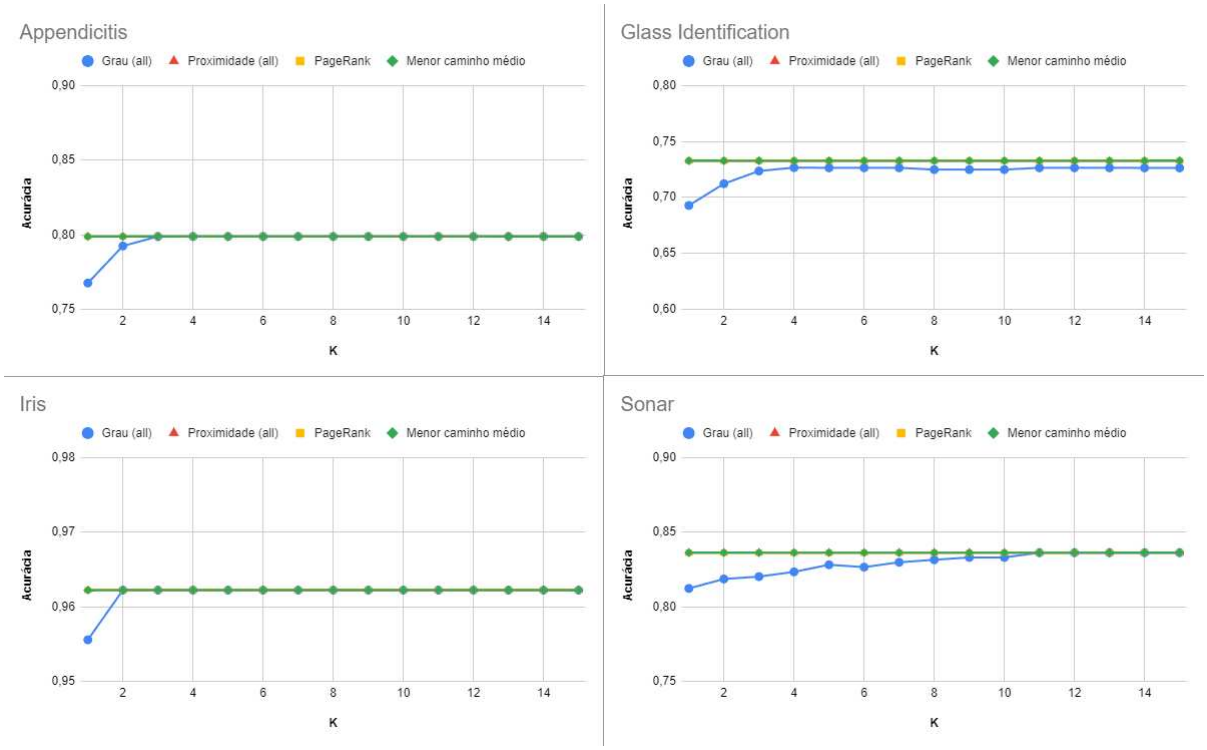


Figura 20 – Variação da acurácia de acordo com o aumento do valor de K com o valor de gamma em 0 para as medidas de redes complexas nas bases reais *Appendicitis*, *Glass identification*, *Iris* e *Sonar*.

A Figura 20 apresenta a variação dos resultados obtidos das medidas de redes com-

plexas para o valor de $\gamma = 0$ em variação do valor K para a construção da rede. Pode-se perceber que com esse valor de γ obtém-se resultados semelhantes para todas as medidas de redes analisadas, mas é possível perceber um menor desempenho da medida de rede grau quando usamos valores mais baixos para K na construção da rede.

A Figura 21 já demonstra uma melhoria no desempenho preditivo de algumas medidas de rede como a medida *PageRank*, grau e proximidade. Mas observa-se que com o aumento do valor de K a medida menor caminho médio sofre uma piora no desempenho na base de dados Iris. Já na base Glass identification a medida de menor caminho médio tem uma melhoria no desempenho, contudo, as demais, em contrapartida, apresentam uma piora em seu desempenho. Já o PageRank na base Appendicitis teve uma queda de desempenho em relação às outras medidas de rede.

Na Figura 22 observa-se que a medida de rede menor caminho médio teve uma piora expressiva nas bases *Apeendicitis*, *Iris* e *Sonar*, enquanto na *Glass identification* se manteve com resultados semelhantes às outras medidas de rede. É possível observar que na base de dados *Sonar* todas as medidas sofreram uma piora no resultado de acurácia. Já na base *Appendicitis* as medidas de rede tiveram uma melhoria no desempenho e a medida *PageRank* se sobressaiu em relação às outras.

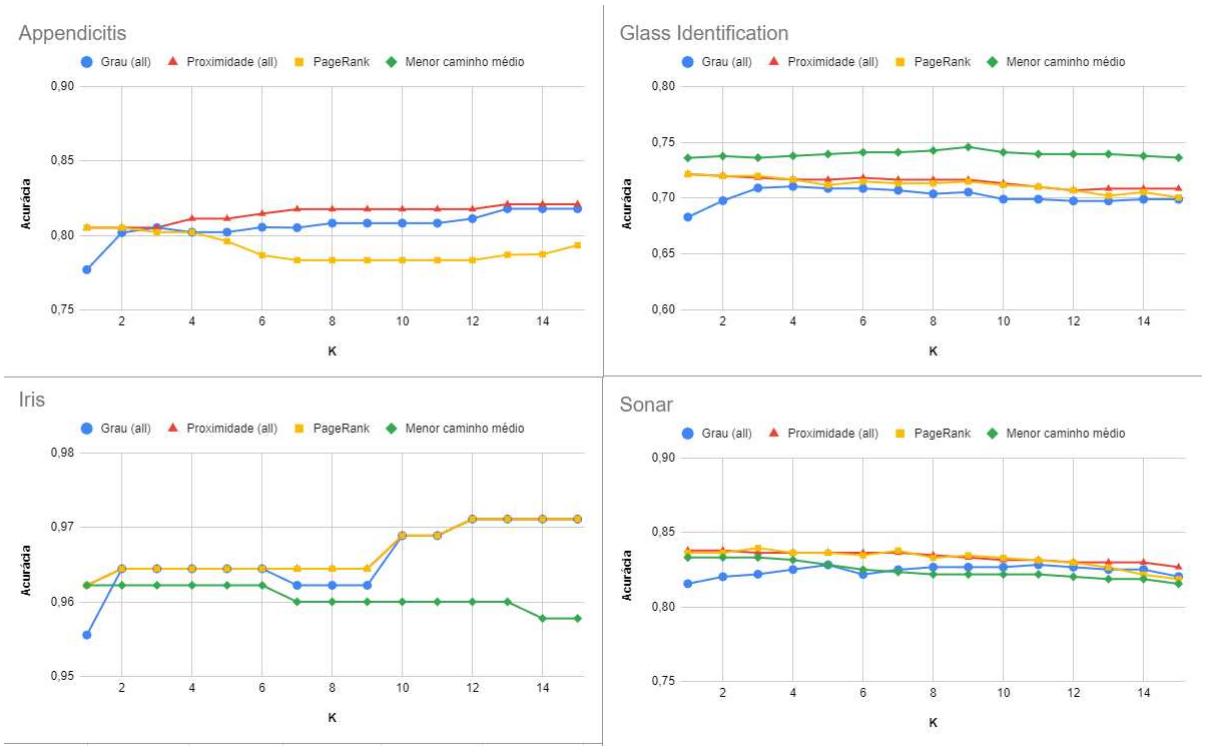


Figura 21 – Variação da acurácia de acordo com o aumento do valor de K com o valor de γ em 0.5 para as medidas de redes complexas nas bases reais *Appendicitis*, *Glass identification*, *Iris* e *Sonar*.

A Figura 23 mostra os resultados obtido pelas medidas de redes complexas em função do valor K para a construção de rede mas com o valor de 2 para o parâmetro γ. É

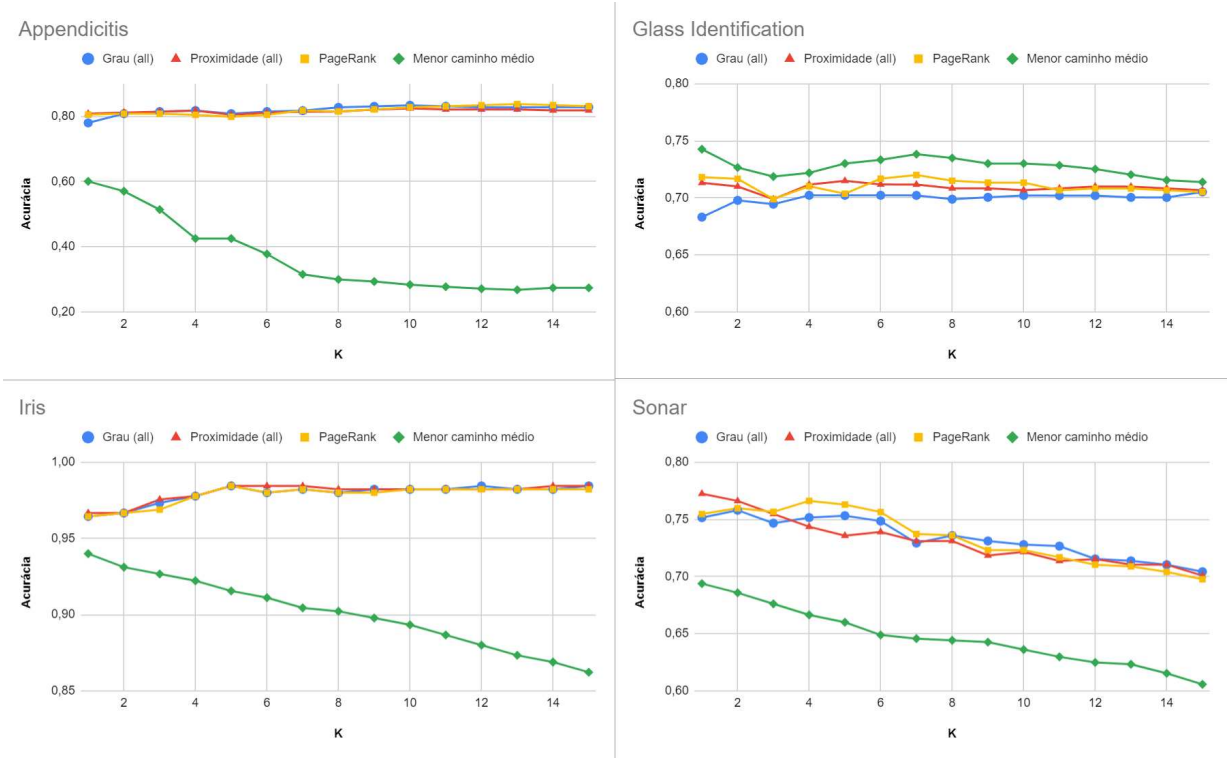


Figura 22 – Variação da acurácia de acordo com o aumento do valor de K com o valor de γ em 1 para as medidas de redes complexas nas bases reais *Appendicitis*, *Glass identification*, *Iris* e *Sonar*.

possível notar a diferença dos resultados e variações das linhas em relação à Figura 20. Podemos notar que na base *Appendicitis* houve uma melhoria nos resultados obtidos para as medidas de redes grau, proximidade e *PageRank*. Em contrapartida a medida de rede menor caminho médio obteve uma piora nos resultados. Já na base *Glass identification* houve uma piora nos resultados para todas as medidas. Na base *Iris* obtém-se uma melhoria de resultado porém mais uma vez a medida de rede menor caminho médio teve uma piora expressiva em relação às outras medidas. E por fim na base *Sonar* também houve uma piora dos resultados e pelo sinal do gráfico a tendência é que a acurácia caia enquanto o valor de K aumenta.

Tabela 5 – Média dos valores de acurácia e desvio padrão das medidas de redes complexos, considerando a variação de ambos os parâmetros γ e K.

Medidas de redes	Appendicitis	Glass	Iris	Sonar
Grau	81±9.0	70±9.0	**97±5.0**	78±7.0
Proximidade	**82±9.0**	71±9.0	**97±5.0**	78±7.0
PageRank	81±10.0	72±9.0	96±5.0	**79±8.0**
Menor Caminho	41±11.0	**73±8.0**	93±5.0	73±7.0

A Tabela 5 apresenta a média de resultados considerando todos os resultados variando o valor de K e o valor de γ. Observa-se que a medida de menor caminho médio teve uma queda considerável na base de dados *Appendicitis* se comparado ao seu melhor resultado

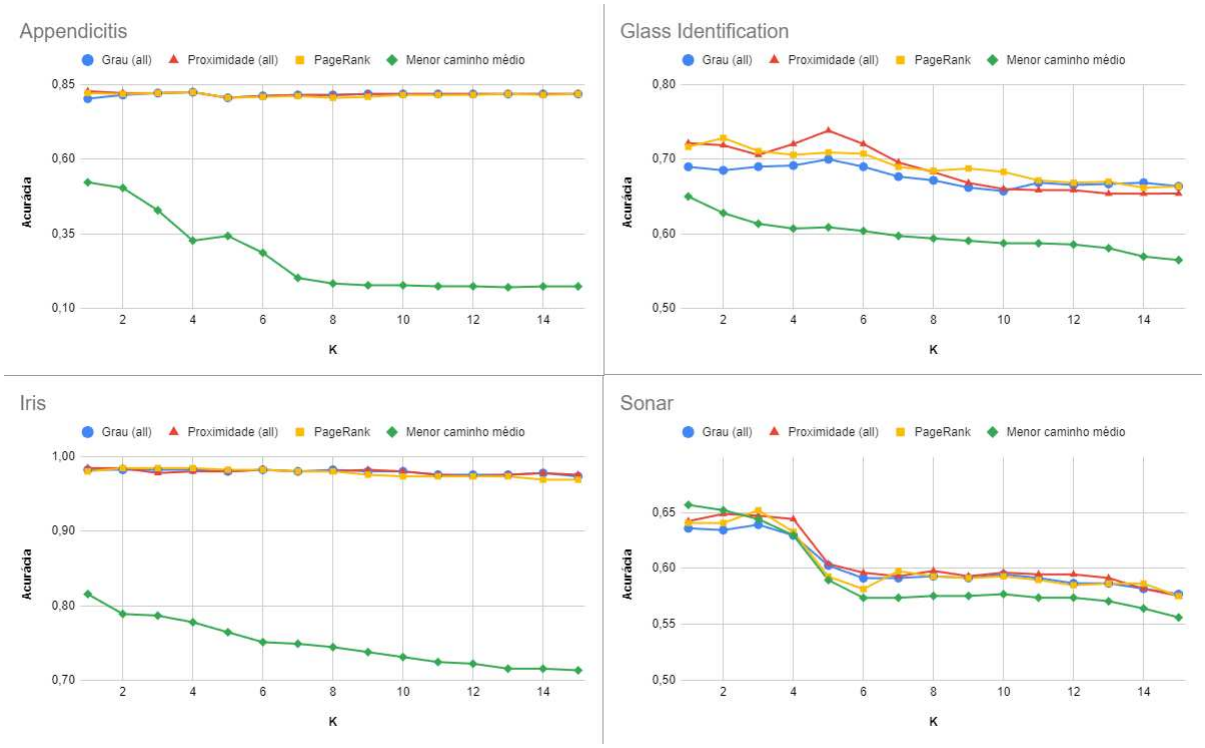


Figura 23 – Variação da acurácia de acordo com o aumento do valor de K com o valor de γ em 2 para as medidas de redes complexas nas bases reais *Appendicitis*, *Glass identification*, *Iris* e *Sonar*.

como mostra a Tabela 4. É possível notar que baseado no resultado das médias a medida de rede proximidade obteve 2 melhores resultados, enquanto que o *PageRank* caiu de 3 melhores resultados no seu melhor cenário para 1 no resultado da média.

4.3 Discussão dos Resultados

Para a avaliação de resultados deve-se levar em conta a sensibilidade à variação do parâmetro K para a construção da rede e à variação do parâmetro γ da técnica de classificação de alto nível.

Quando é analisado apenas a variação do parâmetro K para a construção da rede obtemos resultados bastante lineares em ambos os tipos de base de dados tanto as reais quanto as artificiais, e quando se observa para as bases artificiais com níveis de ruídos maiores o desempenho preditivo das medidas de redes apresentaram uma queda significativa. Portanto olhando para os resultados é possível perceber que para valores baixos do parâmetro K temos um desempenho da medida de rede grau abaixo das demais, mas quando os valores de K são acima de 8 o desempenho aumenta ficando mais semelhante com as demais.

Ao analisar a variação do parâmetro γ, nota-se que as linhas começam a ter mais variações e que nas bases artificiais com nível baixo de ruído como no caso da base de

dados *Moons* há uma queda no desempenho quanto mais se aumenta o parâmetro γ porém quando possui-se a mesma base de dados *Moons* com nível alto de ruído é possível ver uma melhoria no desempenho das medidas apesar de que a medida de rede menor caminho obteve queda de desempenho em todos os níveis de ruído quando se eleva o valor do parâmetro γ. Já na *Classifications* é possível ver que em baixos níveis de ruídos o desempenho de algumas medidas foi parecido. A medida de menor caminho médio acabou demonstrando uma queda nos resultados. Já nas bases reais é possível ver uma queda no desempenho das medidas de redes quando o valor de γ é muito elevado. É possível observar também que algumas medidas de redes acabam sofrendo mais que as outras como nos casos do menor caminho médio. Pode-se extrair destes resultados a efetividade das medidas de redes quando se muda o nível de ruído e o nível de mistura das bases. Quando existe cenários em que os atributos físicos dos dados se destacam para a classificação então observa-se que as técnicas de baixo nível se destacam, portanto valores baixos de γ são mais eficazes. E quando o cenário possui mais ruído e mistura entre as classes, valores mais altos de γ se destacam pois a relevância dos atributos topológicos aumenta.

Com base em todos as análises conduzidas segue algumas considerações:

- A medida de redes complexas menor caminho médio apresentou queda de acurácia para valores maiores de γ. Então observa-se que essa medida é mais favorecida pelas conexões locais.
- Há a possibilidade de um aumento na acurácia com o aumento do valor de γ, mas não são todas as medidas de redes que compartilham desse ganho de acurácia.
- Dados que conseguimos extrair mais informações topológicas conseguem ter melhores resultados de classificação. Resultando em uma melhor capacidade de dedução em níveis maiores de ruído e mistura da base de dados.
- Outras medidas de redes complexas obtiveram resultados de acurácia semelhantes ao do PageRank, o que abre caminho para investigações mais amplas acerca da contribuição de cada medida para a classificação via caracterização de importância.

Capítulo 5

Conclusão

Este trabalho teve como objetivo principal avaliar empiricamente outras medidas de redes complexas diferentes do PageRank utilizando a técnica de classificação baseada em importância proposta em Carneiro e Zhao (2018). A partir de um estudo da literatura voltada à classificação de dados via caracterização de importância, foram selecionadas algumas medidas de redes, a saber: grau, proximidade, *PageRank*, menor caminho médio, para que fosse comparado os resultados em relação ao próprio *PageRank*.

A fim de obter uma análise mais profunda foram selecionadas 2 bases artificiais com 3 níveis diferentes de ruídos e 4 bases reais. Também foi avaliado diferentes valores de parâmetros relacionados à construção da rede e a técnica de classificação via caracterização de importância.

Foram conduzidos vários experimentos para fins de análise e avaliação da eficiência das medidas de redes complexas diante das variações de K e gamma. E a partir dessas duas variações foram levantadas duas análises a fim de observar a sensibilidade das medidas de redes em relação a cada uma das variações.

Na análise de variação de K para a construção da rede, é possível observar que algumas medidas de redes conseguem um desempenho melhor em valores mais altos de K e algumas são prejudicadas com o aumento de K. Por isso é difícil deduzir a eficiência de uma medida de rede apenas por esse parâmetro. Mas fazendo a comparação das outras medidas em relação ao *PageRank*, observa-se uma semelhança de resultados tanto no que se diz crescimento da acurácia quanto na diminuição da acurácia.

Portanto sob a perspectiva da variação de ε, existe que algumas medidas são muito sensíveis a essa variação. Um exemplo dessa sensibilidade pode ser visto nas bases artificiais com a medida de menor caminho médio: quanto mais se aumenta o valor de gamma mais ela diminui o desempenho. Nas bases de dados reais a medida de menor caminho e a de proximidade com pesos nas arestas também demonstraram esse tipo de sensibilidade. Fazendo o comparativo de todas as outras medidas testadas neste trabalho em relação ao *PageRank* também é possível observar uma acurácia semelhante ou até melhor em alguns casos.

5.1 Trabalhos Futuros

- ❑ Considerar um número maior de bases reais e artificiais ampliar as simulações e conduzir testes estatísticos.
- ❑ Utilizar de outros métodos para a construção da rede como por exemplo a rede vizinha de raio ε, ou mesmo combinações entre diferentes heurísticas baseadas no kNN (k-vizinhos mais próximos).
- ❑ Utilizar da literatura outras medidas de redes complexas, visto que as que foram utilizadas no trabalho obtiveram resultados interessantes comparado-as com a medida PageRank

5.2 Produção bibliográfica

As investigações conduzidas nesta dissertação permitiram colaborar no desenvolvimento do artigo (FERNANDES et al., 2023) publicado no IJCNN 2023 (*International Joint Conference on Neural Networks*).

Referências

AGGARWAL, C. C. **Data Classification: Algorithms and Applications**. [S.l.]: Chapman and Hall/CRC, 2014. Citado na página 19.

ALBERT, R.; BARABáSI, A.-L. Statistical mechanics of complex networks. **American Physical Society**, v. 74, n. 1, p. 47 – 97, 2002. Citado na página 21.

BERTINI, J. R. et al. A nonparametric classification method based on k-associated graphs. **Information Sciences**, v. 181, n. 24, p. 5435–5456, 2011. ISSN 0020-0255. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0020025511003823>. Citado 3 vezes nas páginas 9, 26 e 27.

BOLDI, P.; SANTINI, M.; VIGNA, S. Pagerank as a function of the damping factor. In: **international conference on World Wide Web**. [S.l.: s.n.], 2005. p. 557 – 566. Citado na página 29.

CARNEIRO, M. G. et al. Particle swarm optimization for network-based data classification. **Neural Networks**, n. 110, p. 243 – 255, 2019. Citado na página 15.

CARNEIRO, M. G.; GAMA, B. C.; RIBEIRO, O. S. Complex network measures for data classification. In: **International Joint Conference on Neural Networks (IJCNN)**. [S.l.: s.n.], 2021. p. 1–8. Citado na página 30.

CARNEIRO, M. G. et al. High-level classification for EEG analysis. In: **International Joint Conference on Neural Networks (IJCNN)**. [S.l.: s.n.], 2023. p. 1–8. Citado 2 vezes nas páginas 17 e 28.

CARNEIRO, M. G.; ZHAO, L. **Redes complexas para classificação de dados via conformidade de padrão, caracterização de importância e otimização estrutural**. 161 p. Dissertação (Mestrado) — Universidade de São Paulo, São Carlos, 2017. Citado 15 vezes nas páginas 9, 15, 16, 20, 21, 23, 24, 25, 26, 27, 29, 30, 34, 35 e 36.

CARNEIRO, M. G.; ZHAO, L. Organizational data classification based on the importance concept of complex networks. **IEEE Transactions on Neural Networks and Learning Systems**, v. 29, n. 8, p. 3361 – 3373, 2018. Citado 6 vezes nas páginas 15, 16, 17, 28, 34 e 48.

COSTA, L. da F. et al. Characterization of complex networks: A survey of measurements. **Advances in Physics**, Taylor Francis, v. 56, n. 1, p. 167–242, 2007. Disponível em: <https://doi.org/10.1080/00018730601170527>. Citado 2 vezes nas páginas 20 e 21.

FERNANDES, J. M. et al. Data classification via centrality measures of complex networks. In: **International Joint Conference on Neural Networks (IJCNN)**. [S.l.: s.n.], 2023. p. 1–8. Citado na página 49.

FILHO, S. R. A. O.; MOURA, I. V. P. Análise baseada em redes complexas para identificação da estrutura de comando de operação militar como fator de alta vulnerabilidade. **XX Simpósio de Aplicações Operacionais em Áreas de Defesa - SIGE**, p. 5, 9 2018. Citado 2 vezes nas páginas 9 e 21.

FISHER, R. A. **Iris**. 1988. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C56C76. Citado na página 33.

GERMAN, B. **Glass Identification**. 1987. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5WW2P. Citado na página 33.

MOHRI, M.; ROSTAMIZADEH, A.; TALWALKAR, A. **Foundations of machine learning**. [S.l.]: Massachusetts Institute of Technology, 2018. Citado na página 20.

OZAKI, K. et al. Using the mutual k-nearest neighbor graphs for semi-supervised classification of natural language data. **ACL Conference on Computational Natural Language Learning**, p. 154 – 162, 2011. Citado na página 25.

RESENDE, V. H.; CARNEIRO, M. G. Analysis of complex network measures for multi-label classification. **International Journal on Artificial Intelligence Tools**, v. 30, n. 04, p. 2150023, 2021. Disponível em: <https://doi.org/10.1142/S0218213021500238>. Citado 3 vezes nas páginas 15, 20 e 30.

SEJNOWSKI, T.; GORMAN, R. **Connectionist Bench (Sonar, Mines vs. Rocks)**. 1988. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5T01Q. Citado na página 33.

Silva, T. C.; Zhao, L. Network-based high level data classification. **IEEE Transactions on Neural Networks and Learning Systems**, v. 23, n. 6, p. 954–970, June 2012. ISSN 2162-237X. Citado 2 vezes nas páginas 27 e 28.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introdução ao DATAMINING Mineração de Dados**. [S.l.]: Editora Ciência Moderna Ltda., 2009. Citado 2 vezes nas páginas 9 e 19.

ZHU, X.; LAFFERTY, J.; ROSENFELD, R. **Semi-Supervised Learning with Graphs**. 174 p. Dissertação (Mestrado) — Carnegie Mellon University, São Carlos, 2005. Citado na página 25.