
Avaliação comparativa entre medidas de redes complexas para a classificação de dados

Barbara Cristina Gama



UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE COMPUTAÇÃO
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

Monte Carmelo - MG
2020

Barbara Cristina Gama

**Avaliação comparativa entre medidas de redes
complexas para a classificação de dados**

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: Ciência da Computação

Orientador: Dr. Murillo Guimarães Carneiro

Monte Carmelo - MG

2020

Barbara Cristina Gama

Avaliação comparativa entre medidas de redes complexas para a classificação de dados

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.

Monte Carmelo - MG, 16 de janeiro de 2020:

Dr. Murillo Guimarães Carneiro
Orientador

Dr. Carlos Cesar Mansur Tuma
Professor Convidado

Dr. Paulo Henrique Ribeiro Gabriel
Professor Convidado

Monte Carmelo - MG
2020

Dedico a minha mãe Fátima e ao meu pai Valdemar, pela oportunidade de me dedicar unicamente a minha formação, e a todos que acreditaram e fizeram parte deste trabalho, em especial ao Professor Dr. Murillo Carneiro.

Agradecimentos

A Deus por me manter sã, me proteger e tornar possível a conclusão deste trabalho.

Ao meu pai Valdemar, pelo exemplo de honestidade, força e bondade, me motivando todos os dias a seguir seus valores e acreditar que nada é impossível.

A minha mãe Fátima, pelo cuidado, incentivo, sensibilidade e valentia de me ajudar nos momentos ruins, e por celebrar minhas pequenas vitórias.

Ao Gustavo, que me impulsiona sempre a ser melhor, pela compreensão e ajuda nos dias ruins, e por contribuir no meu crescimento durante esses anos juntos.

Ao professor Dr. Murillo Carneiro, que me acompanhou desde o primeiro período da graduação, que me incetiva todos os dias a melhorar, pelo exemplo de dedicação, ética e seriedade com seu trabalho, pela paciência e ensinamentos como professor e orientador, serei eternamente grata por sua orientação e inspiração.

Aos professores Dr. Carlos Cesar Mansur Tuma e Dr. Paulo Henrique Ribeiro Gabriel por aceitarem participar da banca examinadora, neste momento tão importante para mim.

A todos os professores que fizeram parte da minha formação, e aos amigos e colegas que compartilharam desses momentos em todos esses anos.

A todos os envolvidos nos projetos da UFU que participei, onde pude passar algum conhecimento, tornando enriquecedor minha trajetória.

A Universidade Federal de Uberlândia (UFU) e todos os envolvidos pela oportunidade de realizar o curso na minha cidade.

*“Concedei-me Senhor, a serenidade necessária
para aceitar as coisas que não posso modificar,
coragem para modificar aquelas que posso,
e sabedoria para distinguir umas das outras.”*
(Oração da serenidade)

Resumo

Redes complexas são redes com estruturas de conexão não triviais, as quais não seguem um padrão regular nem aleatório. O uso delas para classificação de dados tem se tornado tópico de pesquisa cada vez mais relevante na literatura, com vários trabalhos explorando de maneira bem sucedida medidas e propriedades de redes complexas. Na classificação de dados, o uso de redes complexas possibilita interpretar e modelar as relações semânticas entre os dados, permitindo a detecção de padrões estruturais e topológicos relacionados, por exemplo, à formação de padrão dos dados. Várias medidas de redes complexas já foram utilizadas nesse sentido, porém, falta na literatura um estudo sistemático capaz de caracterizar a influência e o comportamento de tais medidas no contexto da classificação de dados. Como existem muitas medidas de redes complexas que possuem características diferentes, este trabalho consiste em desenvolver um ambiente experimental a fim de analisar e comparar o comportamento de algumas medidas, revelando cenários mais adequados para usar uma ou outra medida. Especificamente, o ambiente desenvolvido contempla a construção da rede a partir da rede k-vizinhos mais próximos, e a exploração de medidas de redes complexas a partir da técnica de classificação baseada em conformidade de padrão (classificação de alto-nível). Para isso, foram selecionadas seis medidas de redes complexas, a saber: assortatividade, coeficiente de agrupamento, grau médio, intermediariedade, menor caminho médio e proximidade. Para as análises, foram consideradas tanto bases artificiais quanto reais, todas na forma de vetor de atributos. Os resultados obtidos apontam que medidas, tais como, menor caminho médio e coeficiente de agrupamento, além de apresentarem melhor desempenho preditivo nos cenários com maior nível de ruído, também são mais robustas à variação do parâmetro de construção da rede. Em suma, esta pesquisa consiste em uma importante iniciativa para auxiliar outros trabalhos relacionados na seleção das medidas de redes complexas para classificação de dados.

Palavras-chave: Aprendizado de máquina, redes complexas, classificação de dados, medidas de redes complexas, conformidade de padrão.

Abstract

Complex networks are networks with nontrivial connection structures, which do not follow a regular or random pattern. Their usage for data classification has become an increasingly relevant research topic in recent years, with many techniques in the literature successfully exploring complex network properties and measures. In data classification, the use of complex networks allows to interpret and model semantic relationships among data items, enabling the detection of structural and topological patterns related, for example, to the formation pattern of the input data. Several measures of complex networks have already been used in this sense. However, a systematic study capable of characterizing the behavior and performance of such measures in the context of data classification is lacking in the literature. As there are many measures of complex networks that have different characteristics, this work consists of developing an experimental environment in order to analyze and compare the behavior of some measures, revealing more adequate measures for one or another circumstance. Specifically, the computational environment involves to build up a network by applying the k-nearest neighbor criterion over the input data, and to explore complex network measures through of the classification via pattern conformation (a.k.a. high-level classification). For this purpose, six complex network measures were selected from the literature, namely: assortativity, clustering coefficient, average degree, betweenness, shortest average path and closeness. For our analyses, both artificial and real-world data sets represented as feature vectors were considered. The results show that measures such as shortest average path and clustering coefficient, besides presenting better predictive performance in data sets with higher noise level, are also more robust to the variation of the network construction parameter. In summary, this research is an important initiative to support other related works in selecting more appropriate complex network measures for data classification.

Keywords: Classificação de alto-nível, machine learning, complex networks, data classification, complex network measures.

Lista de ilustrações

Figura 1 – Exemplo de problema real estudado em redes complexas - Internet representada como uma rede (NEWMAN, 2003).	14
Figura 2 – Exemplo de problema real estudado em redes complexas - Rede “Caratê” (ZACHARY, 1977).	15
Figura 3 – Exemplo de classificação de dados, utilizando uma galáxia espiral e uma galáxia elíptica. Imagens obtidas no website da NASA.	18
Figura 4 – Exemplo do algoritmo KNN para classificação de dados, retirado de Ferrero (2009).	19
Figura 5 – Exemplo de grafo completo com cinco nós.	21
Figura 6 – Exemplo do funcionamento da conformidade padrão na classificação de dados, retirado de Silva e Zhao (2012).	23
Figura 7 – Bases artificiais geradas a partir do domínio Moons, considerando três níveis de ruídos, denotados como baixo, médio e alto	27
Figura 8 – Bases artificiais geradas a partir do domínio Circle, considerando três níveis de ruídos, denotados como baixo, médio e alto	27
Figura 9 – Bases artificiais geradas a partir do domínio Classification, considerando três níveis de ruídos, denotados como baixo, médio e alto	28
Figura 10 – Variação da acurácia das medidas de redes complexas para as redes do domínio Moons, e cada uma indicando o nível de ruído (baixo, médio e alto), em função de k	32
Figura 11 – Variação da acurácia das medidas de redes complexas para as redes do domínio Circles, e cada uma indicando o nível de ruído (baixo, médio e alto), e mostra qual medida de redes complexas tem melhor resultado para as bases.	35
Figura 12 – Variação da acurácia das medidas de redes complexas para as redes do domínio Classification, e cada uma indicando o nível de ruído (baixo, médio e alto), e mostra qual medida de redes complexas tem melhor resultado para as bases.	37

Figura 13 – Variação da acurácia das medidas de redes complexas para as bases reais (Appendicitis, Balance, Iris e Sonar).	38
---	----

Lista de tabelas

Tabela 1	– Descrição das propriedades das bases reais utilizadas no trabalho (Appendicitis, Balance, Iris e Sonar), para cada base tem-se o número de instâncias, a quantidade de atributos e quantas classes.	28
Tabela 2	– Melhores valores de acurácia e desvio padrão das medidas de redes complexas para o domínio Moons. Entre parênteses é apresentado o valor de K usado para construir a rede.	32
Tabela 3	– Acurácia e desvio padrão das medidas de redes complexas para o domínio Moons, considerando todas as variações do parâmetro de formação da rede (k).	33
Tabela 4	– Melhores valores de acurácia e desvio padrão das medidas de redes complexas para o domínio Circles. Entre parênteses é apresentado o valor de K usado para construir a rede.	34
Tabela 5	– Acurácia e desvio padrão das medidas de redes complexas para o domínio Circles, considerando todas as variações do parâmetro de formação da rede (k).	34
Tabela 6	– Melhores valores de acurácia e desvio padrão das medidas de redes complexas para o domínio Classification. Entre parênteses é apresentado o valor de K usado para construir a rede.	36
Tabela 7	– Acurácia e desvio padrão das medidas de redes complexas para o domínio Classification, considerando todas as variações do parâmetro de formação da rede (k).	36
Tabela 8	– Melhores valores de acurácia e desvio padrão das medidas de redes complexas para para as bases reais (Appendicitis, Balance, Iris e Sonar). Entre parênteses é apresentado o valor de K usado para construir a rede.	37
Tabela 9	– Acurácia e desvio padrão das medidas de redes complexas para as bases reais (Appendicitis, Balance, Iris e Sonar), considerando todas as variações do parâmetro de formação da rede (k).	39

Sumário

1	INTRODUÇÃO	13
1.1	Objetivos	15
1.2	Hipótese	16
1.3	Organização da Monografia	16
2	FUNDAMENTAÇÃO TEÓRICA	17
2.1	Classificação de dados	17
2.1.1	Classificador K-vizinhos mais próximos (KNN)	18
2.1.2	Medidas de avaliação	19
2.2	Medidas de Redes Complexas	20
2.2.1	Grau Médio	20
2.2.2	Assortatividade	20
2.2.3	Coefficiente de Agrupamento	21
2.2.4	Intermedialidade	21
2.2.5	Menor Caminho Médio	22
2.2.6	Proximidade	22
2.3	Classificação de dados em redes complexas	22
2.3.1	Grafo K-vizinhos mais próximos (KNN)	23
2.3.2	Conformidade Padrão	23
2.4	Trabalhos relacionados	24
3	MATERIAIS E MÉTODOS	26
3.1	Bases de dados	26
3.1.1	Bases artificiais	26
3.1.2	Bases reais	27
3.2	Algoritmo para avaliação das medidas de rede	29
3.3	Ambiente experimental	30

4	RESULTADOS EXPERIMENTAIS	31
4.1	Resultados em bases artificiais	31
4.2	Resultados em bases reais	36
4.3	Discussão dos resultados	39
5	CONCLUSÃO	41
	REFERÊNCIAS	43

Introdução

Nos dias atuais, a tecnologia proporciona uma facilidade para a criação de um acervo cada vez maior e diversificado de dados. O aprendizado de máquina tem como objetivo explorar tais coleções de dados, com a finalidade de “aprender” padrões e correlações existentes nelas. Dentre as várias tarefas do aprendizado de máquina, a classificação de dados é considerada uma das mais comuns. Um exemplo clássico desta tarefa é a classificação de *spam*, onde a partir de um grupo de *emails* previamente conhecidos como *spam* ou não, treina-se um algoritmo para classificar novos *emails* (CARNEIRO, 2016).

Apesar da classificação de dados ser uma tarefa bastante investigada na literatura, estudos recentes (SILVA; ZHAO, 2012; CUPERTINO et al., 2018; CARNEIRO; ZHAO, 2018; CARNEIRO et al., 2019) demonstraram que as principais técnicas de classificação, tais como árvore de decisão, redes neurais, máquinas de vetores de suporte, etc., possuem dificuldade em capturar padrões semânticos dos dados, tais como a formação de padrão, uma vez que são baseadas apenas nos atributos físicos dos dados (distância, similaridade ou distribuição, por exemplo). Por outro lado, a análise de propriedades estruturais e topológicas dos dados através de medidas de redes complexas tem se mostrado eficiente nesse propósito.

Muitos sistemas complexos são representados por redes atualmente. Um exemplo bastante conhecido de estrutura modelada por rede é a Internet, cuja rede é apresentada na Figura 1. De modo geral, redes representam relacionamentos, denominados arestas, entre pares de objetos de qualquer tipo, denominados vértices, com vários *insights* sendo possíveis a partir da análise dessas relações e dos padrões de conexão existentes. Com tais características, redes complexas são utilizadas para entender como e por que os objetos se conectam e qual o impacto destas conexões (LOPES, 2011).

Um exemplo clássico de análise dos padrões de conexão pode ser visto na rede “Caratê”, apresentada na Figura 2. Ela foi compilada em Zachary (1977) e apresenta as relações entre os membros de uma academia de caratê, incluindo o instrutor e o proprietário, tomando como referência a relação entre os membros durante eventos sociais da academia. Após alguns desentendimentos entre eles, o instrutor foi então demitido

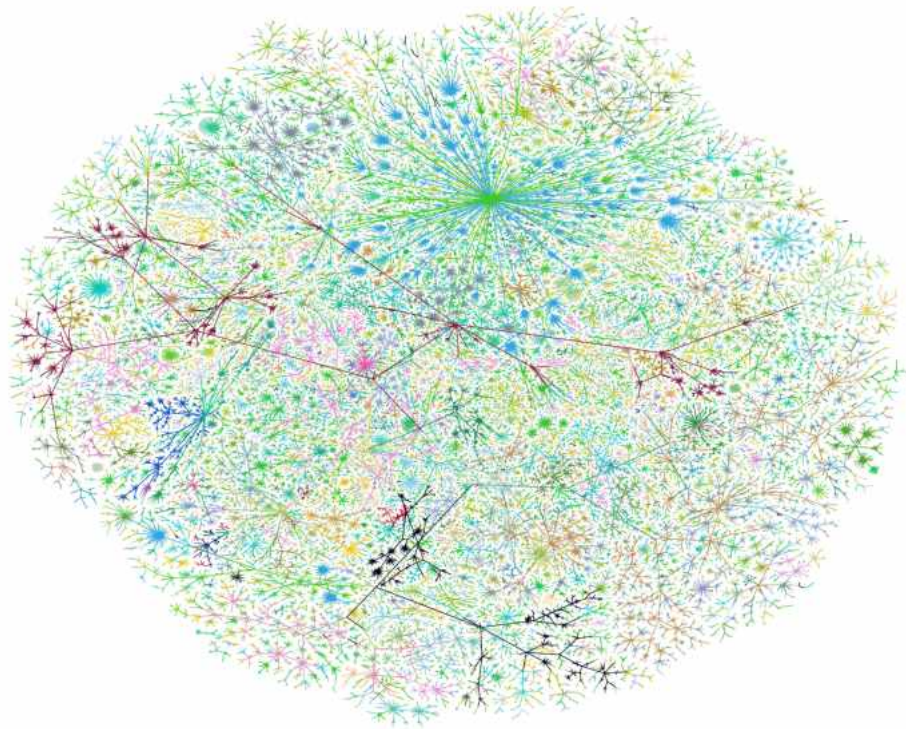


Figura 1 – Exemplo de problema real estudado em redes complexas - Internet representada como uma rede (NEWMAN, 2003).

da academia, o que resultou no desligamento de vários alunos. Tal fenômeno pode ser entendido a partir da análise da própria rede. Os vértices 1 e 34 (na cor vermelha) denotam respectivamente o instrutor e o proprietário. Os membros da academia diretamente relacionados com o vértice 1 estão na cor amarela, enquanto os de cor verde representam membros com relacionamento direto com o vértice 34. Na cor azul estão os alunos que tem relacionamento direto com os dois instrutores, e os que estão em branco não tem ligação direta, porém estão indiretamente ligados a algum deles por meio de outros alunos. Na figura, é fácil perceber que um grupo de membros possui forte relação com o instrutor, enquanto outro grupo está bem mais relacionado com o proprietário.

A classificação de dados baseada em redes complexas tem sido alvo de estudos recentes relacionados especialmente à construção da rede (CARNEIRO; ZHAO, 2018; CARNEIRO et al., 2019) e à exploração de propriedades e medidas de redes complexas (SILVA; ZHAO, 2012; CUPERTINO et al., 2018; CARNEIRO; ZHAO, 2018). Na construção da rede, novas abordagens têm sido desenvolvidas para substituir abordagens convencionais, tais como a rede k-vizinhos mais próximos. Na exploração das medidas, novos conceitos de classificação de dados, tais como conformidade padrão e caracterização de importância, têm sido desenvolvidos a partir de medidas de redes complexas. A conformidade padrão, ou classificação de alto nível segundo Silva e Zhao (2012), realiza a classificação de um novo objeto por verificar a conformidade de padrão do mesmo em relação aos subcomponentes

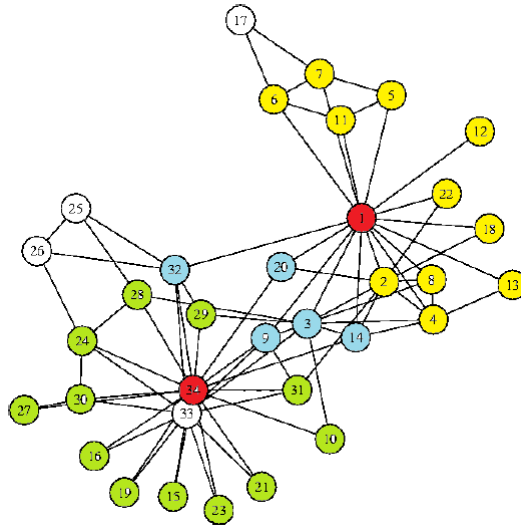


Figura 2 – Exemplo de problema real estudado em redes complexas - Rede “Caratê” (ZACHARY, 1977).

da rede (classes), assim, um novo objeto é classificado no subcomponente em que a sua inserção causa a menor variação de um conjunto de medidas de redes complexas definido a priori. Já na caracterização de importância, ou classificação baseada em importância segundo Carneiro e Zhao (2018), a classificação de um novo objeto se dá em função de um valor de importância que ele recebe quando inserido nos subcomponentes da rede, de modo que o objetivo é associar o objeto à classe na qual recebe maior importância.

Como o uso de redes complexas ganha cada vez mais relevância em aplicações de aprendizado de máquina supervisionado, uma pesquisa empírica sobre as contribuições de um conjunto relevante de medidas para a classificação de dados é um trabalho necessário na área, visto que não há estudos para caracterizar medidas de redes complexas em classificação de dados. Como existem muitas medidas de redes complexas que possuem características muito diferentes, a finalidade do trabalho é realizar a comparação entre elas, diante de vários problemas reais e artificiais, revelando cenários mais adequados para considerar uma ou outra medida.

1.1 Objetivos

Este trabalho tem como objetivo principal avaliar comparativamente um conjunto de medidas de redes complexas para a classificação de dados. Especificamente, os seguintes objetivos são pontuados:

- Selecionar um conjunto de medidas de redes complexas relevantes no contexto da classificação de dados.

- ❑ Adaptar o algoritmo de classificação via conformidade de padrão para permitir a análise comparativa das medidas de redes complexas individualmente.
- ❑ Projetar e desenvolver um ambiente experimental propício para caracterizar empiricamente as medidas de redes tanto em cenários artificiais com diferentes níveis de ruído quanto em bases reais de diferentes domínios.

1.2 Hipótese

A hipótese investigada neste trabalho afirma que há cenários específicos da classificação de dados onde certas medidas de redes complexas podem apresentar melhor desempenho preditivo (acurácia) do que outras.

1.3 Organização da Monografia

O restante desta dissertação é dividido nos seguintes capítulos:

- ❑ O Capítulo 2 apresenta conceitos fundamentais para o entendimento e desenvolvimento do trabalho: classificação de dados, redes complexas, classificação de dados em redes complexas, conformidade padrão, e os principais trabalhos relacionados da literatura.
- ❑ O Capítulo 3 descreve todos os materiais e métodos empregados para realizar este trabalho, incluindo a coleta e tratamento das bases de dados, construção das redes utilizando o algoritmo KNN (K-vizinhos mais próximos), apresentação do algoritmo de classificação via conformidade padrão adotado e do ambiente computacional desenvolvido para os experimentos.
- ❑ O Capítulo 4 discute os resultados obtidos pelas medidas de redes tanto nas bases artificiais quanto reais em termos de desempenho preditivo e sensibilidade à variação do parâmetro.
- ❑ O Capítulo 5 apresenta as conclusões e perspectivas futuras para o trabalho.

Fundamentação Teórica

Neste capítulo é apresentada uma revisão da literatura com os principais conceitos e assuntos para o entendimento e realização do trabalho. Na Seção 2.1 é introduzida a classificação de dados, incluindo o algoritmo de classificação K-vizinhos mais próximos (KNN) e as principais métricas de desempenho utilizadas para a tarefa. Na Seção 2.2, introduz-se redes complexas com foco específico nas medidas de redes complexas utilizadas neste trabalho, a saber: assortatividade, coeficiente de agrupamento, grau médio, intermedialidade, menor caminho médio e proximidade. Na Seção 2.3 discute-se a classificação de dados baseada em redes complexas, incluindo aspectos da formação da rede utilizando o grafo k-vizinhos mais próximos, e as técnicas de classificação via conformidade padrão e caracterização de importância. Por fim, a Seção 2.4 traz os principais trabalhos relacionados encontrados na literatura.

2.1 Classificação de dados

A classificação de dados é uma das principais tarefas do aprendizado de máquina. Segundo Tan, Steinbach e Kumar (2009), a classificação consiste em aprender um modelo a partir de um conjunto de dados de treinamento, onde cada objeto é denotado pelo par (X, Y) , em que X denota os seus atributos, e Y a sua classe. Logo, o objetivo é que o modelo consiga mapear os atributos dos dados em função das classes, de modo que possa classificar novos objetos de forma eficiente. Assim, define-se que classificação é a tarefa de aprender uma função alvo F – modelo de classificação – que mapeie um conjunto de atributos X para cada classe Y , representada nos dados de treinamento.

Um exemplo simples de classificação de dados é a classificação de galáxias baseada em seus formatos, apresentado na Figura 3. Dado o conceito de classificação acima, e o exemplo da Figura 3, é possível classificar as galáxias em duas classes (Y): galáxia espiral e galáxia elíptica. Para obter essa classificação é necessário extrair um conjunto de atributos (X), como por exemplo a densidade e a forma das galáxias, e aprender um modelo de classificação a partir dele. Para exemplificar uma classificação de dados, se

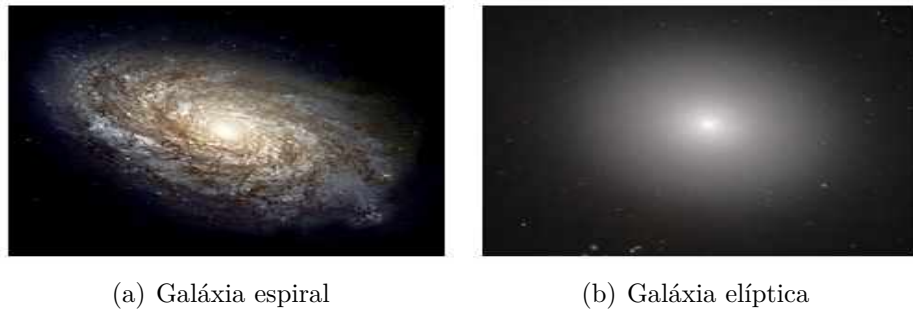


Figura 3 – Exemplo de classificação de dados, utilizando uma galáxia espiral e uma galáxia elíptica. Imagens obtidas no website da NASA.

analisarmos apenas o atributo da forma das galáxias, temos que:

- Se um objeto apresentar forma esférica ou elipsoidal, pode-se definir que o objeto é uma galáxia elíptica - Figura 3(b).
- Se o objeto apresentar forma espiral em volta de um núcleo, pode-se dizer que o objeto é uma galáxia espiral - Figura 3(a).

2.1.1 Classificador K-vizinhos mais próximos (KNN)

O classificador K-vizinhos mais próximos é um dos algoritmos de classificação mais conhecidos na literatura. Na etapa de treinamento, o algoritmo basicamente armazena o conjunto de dados na memória. Na etapa de teste, para classificar um novo objeto, o algoritmo calcula a proximidade desse objeto para todos os outros do conjunto de treinamento e atribui ao objeto a classe majoritária dos seus k vizinhos mais próximos, sendo k um parâmetro do algoritmo.

A Figura 4 apresenta uma aplicação do algoritmo para um conjunto de dados, onde os objetos com rótulo positivo (+) referem-se a pacientes doentes, e objetos com rótulos negativos (-) referem-se a pacientes saudáveis. Para classificação de um objeto/paciente (Ei), a figura ilustra o resultado do classificador KNN para dois valores de k (1 e 4). Dada uma medida de similaridade, e considerando $k=1$, ou seja 1NN, esse novo objeto é classificado como positivo (+), o paciente está doente (Figura 4(a)). Quando $k=4$, o novo objeto será classificado como negativo (-), o paciente não está doente (Figura 4(b)).

A medida de proximidade normalmente utilizada com o KNN é a distância Euclidiana:

$$D = \sqrt{\sum_{k=1}^n (p_{ik} - p_{jk})^2} \quad (1)$$

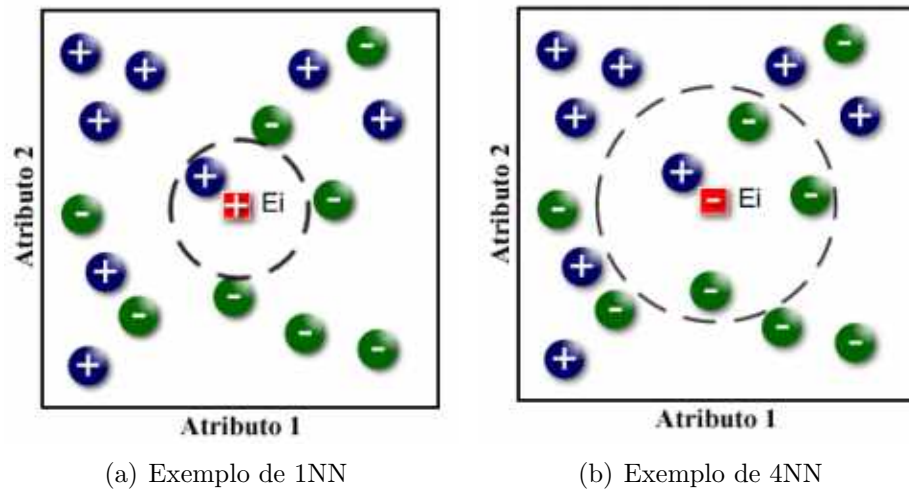


Figura 4 – Exemplo do algoritmo KNN para classificação de dados, retirado de Ferrero (2009).

2.1.2 Medidas de avaliação

Nessa seção serão apresentadas as medidas de avaliação que utilizamos no trabalho, a acurácia e o desvio padrão.

□ Acurácia

A acurácia é a capacidade de classificar corretamente novos objetos em uma base de dados. Essa medida é baseada nas contagens de registros de testes previstos corretamente e incorretamente pelo modelo, ou seja, a acurácia mede a porcentagem dos objetos dos dados de teste que são corretamente classificados. Dado um conjunto X de dados de teste, e C o número de objetos classificados corretamente, temos a Equação 2 da acurácia.

$$Ac = \frac{C}{|X|} 100 \quad (2)$$

□ Desvio padrão

O desvio padrão é uma medida que expressa o grau de dispersão de um conjunto de dados, ou seja, indica o quanto um conjunto de dados é semelhante. Quanto mais próximo de 0 for o desvio padrão, mais uniforme são os dados. A Equação 3 define o desvio padrão, onde X_i é o valor do objeto i , M_a a média aritmética dos dados, e n o número de dados.

$$Dp = \sqrt{\frac{\sum_{i=1}^n (X_i - M_a)^2}{n}} \quad (3)$$

2.2 Medidas de Redes Complexas

Redes complexas contemplam um conjunto de ferramentas para modelagem e análise de sistemas reais. Uma vez que a representação em rede de tais sistemas possuem características topológicas não triviais, isto é, nem completamente regulares nem completamente aleatórias, elas são denominadas redes complexas (CARNEIRO, 2016).

Uma rede representa uma relação entre um ou mais objetos. Para a representação e modelagem de uma rede são utilizados conceitos da teoria dos grafos. Um grafo ou rede é definido como um conjunto de vértices (nós) e um conjunto de arestas que interliga estes nós. Sendo assim, a representação em rede modela as relações existentes em um conjunto de objetos a partir de vértices e arestas.

A literatura apresenta um conjunto considerável de medidas de redes complexas utilizadas em várias aplicações (NEWMAN, 2003; COSTA et al., 2007). Para este estudo, foram selecionadas seis medidas de redes tomando por base artigos relacionados a redes complexas e classificação de dados, além do contexto de grafos não direcionados. Se as arestas apresentam origem e destino o grafo é dito direcionado, se não ele é denominado não direcionado. A seguir são discutidas cada uma das medidas consideradas.

2.2.1 Grau Médio

O grau é uma medida simples, que define o número de conexões de um determinado vértice da rede. O grau de um vértice é representado por:

$$K_i = \sum_{j=1}^n a_{ij} . \quad (4)$$

O grau médio é uma medida que define a quantidade média de vértices adjacentes considerando todos os vértices em uma rede. O grau médio de uma rede é representado por:

$$K = \frac{1}{n} \sum_{i=1}^n K_i \quad (5)$$

2.2.2 Assortatividade

A medida de assortatividade mede a propensão de vértices se conectarem a outros vértices considerando características em comum, ou seja, a medida analisa se os vértices dessa conexão se comportam de forma similar ou não, verificando os graus dos vértices. O coeficiente de assortatividade r , é o coeficiente de correlação de Person entre pares de nós conectados, e pode assumir valores entre $[-1, 1]$, de forma que valores positivos representam maior probabilidade dos vértices se comportarem de forma similar e, valores negativos maior probabilidade de que os vértices estejam conectados de forma distinta (NEWMAN, 2003).

A assortatividade é dada pela Equação 6, onde L é o número de conexões na rede e i_u , k_u os graus dos vértices i e k , que compõe uma aresta u . Assim, é possível determinar, por exemplo, se vértices que tem um grau elevado estão conectados com vértices que também possuam grau elevado.

$$R = \frac{L^{-1} \sum_u i_u - [L^{-1} \sum_u \frac{1}{2}(i_u + k_u)]^2}{L^{-1} \sum_u \frac{1}{2}(i_u^2 + k_u^2) - [L^{-1} \sum_u \frac{1}{2}(i_u + k_u)]^2} \quad (6)$$

2.2.3 Coeficiente de Agrupamento

O coeficiente de agrupamento mede a tendência que os vértices em uma rede possuem para se agrupar (CARNEIRO, 2016), o que indica a probabilidade de que dois vizinhos tenham um outro vizinho em comum. Assim, ele quantifica quão próximo uma rede está de um clique (grafo completo). Um exemplo de clique está representado na Figura 5, onde todos os vértices se conectam com todos os outros e o coeficiente de agrupamento de cada um deles é igual a 1.

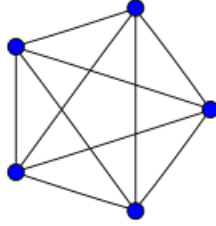


Figura 5 – Exemplo de grafo completo com cinco nós.

O coeficiente de agrupamento é representado pela Equação 7. Supondo que o vértice i possua k vizinhos, há possibilidade de no máximo $k_i(k_i - 1)$ arestas entre eles caso qualquer vizinho u de i é conectado a qualquer outro vizinho s de i .

$$CC_i = \frac{|e_{us}|}{K_i(K_i - 1)} \quad (7)$$

Onde e_{us} é o número de conexões dos vizinhos diretos do vértice i e k_i é o grau do vértice i . Assim, o coeficiente de agrupamento médio da rede é obtido pela Equação 8.

$$CC = \frac{1}{V} \sum_{i=1}^V CC_i \quad (8)$$

2.2.4 Intermedialidade

A intermedialidade, também conhecida como *betweenness* é uma medida de centralidade de vértices, assim, os vértices com maiores valores associados a essa medida, são mais acessados por caminhos mínimos (BARBIERI, 2010).

Em Carneiro (2016) o grau de intermediação B de um vértice i equivale ao número de caminhos geodésicos η (menor caminho da rede), que possuem i no percurso, ou seja,

avalia-se a presença do vértice i no menor caminho existente de um vértice u para outro vértice v , tal que $u, v \in V - i$. Para o vértice i a medida é definida como:

$$B = \sum_{u,v \in V-i} \eta^i_{uv} \quad (9)$$

Segundo Amancio (2013), vértices com altos valores de intermedialidade possuem considerável influência na rede devido ao controle da distribuição de informação. Em sistemas de informação, tais vértices são fundamentais para manter a estrutura da rede, pois sua remoção causa grande impacto na eficiência de distribuição de informação. Uma grande desvantagem da medida é o alto custo computacional dela.

2.2.5 Menor Caminho Médio

Em Vera (2011) define-se que o menor caminho médio ou distância geodésia entre par de nós (i,j) é o menor caminho dentre todos os possíveis caminhos que conectam esses nós. A importância da medida ressalta-se quanto ao transporte de informações e comunicação dentro de uma rede, de forma que o menor caminho permite uma transferência rápida e econômica.

Considere um grafo direcionado não ponderado G com o conjunto de vértices V . $D(i,j)$, onde i e $j \in V$ indica a menor distância entre i e j . Então, o caminho médio MC é dado pela Equação 10, onde n é o número de vértices.

$$MC = \frac{1}{n(n-1)} \sum_{i \neq j} D(i,j) \quad (10)$$

2.2.6 Proximidade

A proximidade mede quanto um vértice estão próximos geograficamente de todos os outros vértices da rede (CARNEIRO, 2016). A proximidade média é dada pela Equação 11, onde $d(i, j)$ é o caminho mínimo entre os vértices, e n é o número de vértices da rede.

$$P = \frac{1}{n} \sum_{i=1}^n \left(\frac{n-1}{\sum_{j=1}^n d(i,j)} \right) \quad (11)$$

2.3 Classificação de dados em redes complexas

Embora o uso de redes complexas em classificação de dados seja recente na literatura, os resultados obtidos até o momento são promissores. Como outros algoritmos de classificação em geral, os algoritmos existentes podem ser divididos em duas maiores etapas: treinamento e teste. Na etapa de treinamento, a rede precisa ser formada caso os dados não estejam na forma de um grafo. Vale ressaltar que a maioria das bases de dados estão representadas na forma de vetor de atributos (CARNEIRO; ZHAO, 2018). Na

etapa de teste, medidas e propriedades de redes complexas são exploradas para realizar a classificação.

2.3.1 Grafo K-vizinhos mais próximos (KNN)

Para a construção da rede a partir de dados do tipo vetor de atributos, um dos métodos mais conhecidos é o grafo K-vizinhos mais próximos. Nesse método cada item de dado é representado por um vértice e haverá uma conexão de um vértice V_i para outro V_j se e somente se V_j pertence ao conjunto dos K elementos mais similares à V_i e ambos V_i e V_j possuem a mesma classe. Para a verificação das relações de vizinhança o grafo KNN utiliza uma medida de proximidade, normalmente a distância Euclidiana (CARNEIRO; ZHAO, 2018).

2.3.2 Conformidade Padrão

A classificação via conformidade de padrão foi originalmente proposta em Silva e Zhao (2012). Em poucas palavras, este algoritmo utiliza um conjunto de medidas de redes complexas para calcular o impacto causado pela inserção de um objeto de teste nos componentes (classes) da rede. O novo objeto é atribuído para a classe em que sua inserção causa a menor variação nas medidas de redes complexas, ou seja, na qual ele apresenta maior conformidade padrão (CARNEIRO, 2016).

Tomada de Silva e Zhao (2012), a Figura 6 (a) apresenta uma rede artificial com três componentes (classes). Para cada componente, as medidas de redes complexas são calculadas. Na Figura 6 (b), a conformidade padrão é utilizada para a classificação de um novo objeto (triângulo rosa). Após a inserção virtual do novo objeto nos componentes da rede, as medidas de rede são recalculadas, obtendo-se assim a variação causada pelo novo objeto em cada componente da rede.

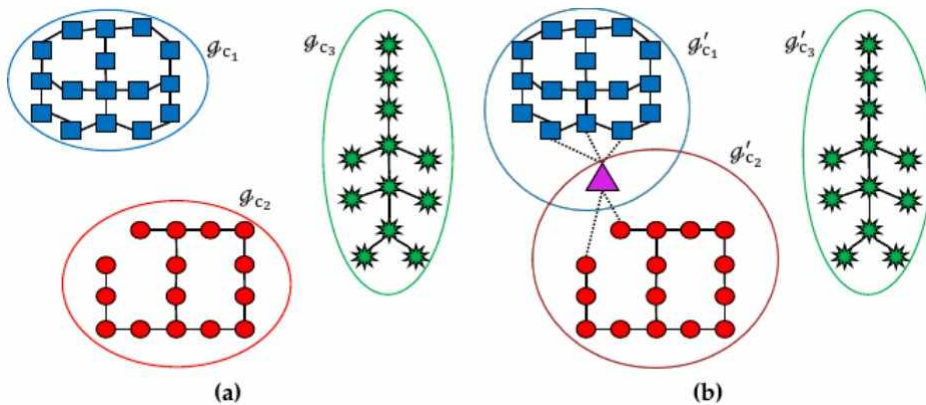


Figura 6 – Exemplo do funcionamento da conformidade padrão na classificação de dados, retirado de Silva e Zhao (2012).

2.4 Trabalhos relacionados

Como afirmado na seção anterior, a classificação de dados baseada em redes complexas normalmente pode ser dividida em duas grandes etapas: treino e teste. Na etapa de treino, o mais comum é a representação dos dados na forma de vetor de atributos na forma de rede, denominada construção da rede (ou grafo). Na etapa de teste, várias abordagens podem ser utilizadas a fim de tirar vantagem de propriedades estruturais, dinâmicas e topológicas dos dados.

Na construção da rede, a maioria das abordagens existentes são baseadas no grafo k-vizinhos mais próximos (Bertini Jr., 2010; CARNEIRO; ZHAO, 2018), embora também seja comum o uso de métodos baseados no grafo de vizinhança- ϵ (ARAÚJO, 2015), ou mesmo a combinação de ambos (SILVA; ZHAO, 2012). Por outro lado, outros trabalhos tem focado na construção de grafos otimizados, os quais permitem obter melhor desempenho (CARNEIRO et al., 2016; CARNEIRO et al., 2017; CARNEIRO et al., 2019), embora apresentem maior custo computacional.

Relacionado à exploração das medidas de rede, além da classificação via conformidade de padrão explicada na seção anterior, outras abordagens se destacam na literatura. Dado isso, temos alguns trabalhos relacionados:

- ❑ Em Bertini Jr. (2010), foi desenvolvido dois algoritmos (KAOG e KAOGINC) baseado em grafo K-associado ótimo e na medida de pureza, com o objetivo de resolver problemas de dados com distribuição estacionária e não estacionária. O primeiro algoritmo demonstrou um bom desempenho com o uso de grafos demonstrando um potencial em problemas de classificação, principalmente por não possuir parâmetros e não exigir seleção de modelos.
- ❑ Em Carneiro et al. (2014), o trabalho tem como objetivo a combinação do grafo K-associados ótimo com a classificação via conformidade de padrão. O método tenta juntar as vantagens dos dois algoritmos: poucos parâmetros e possibilidade de considerar o padrão de formação dos dados.
- ❑ Em Cupertino et al. (2018), foi desenvolvido uma técnica de classificação via conformidade de padrão que usa tanto a variação das medidas de rede quanto uma caminhada aleatória para realizar a classificação.
- ❑ Em Carneiro e Zhao (2018), o trabalho tem como objetivo o desenvolvimento de um método de classificação baseado em importância. Basicamente, o algoritmo realiza a classificação por considerar a importância de cada objeto da rede, assim utilizando as medidas de redes complexas de forma a classificar um novo objeto na classe em que ele recebe maior importância. Para atribuir a importância de cada objeto, o algoritmo é baseado no PageRank, método utilizado pelo Google para caracterizar a importância de páginas Web.

Em técnicas convencionais, a classificação de dados é realizada considerando os atributos físicos dos dados, como a distância, por exemplo. Diferentemente disso, a classificação baseada em redes complexas tem como objetivo capturar a estrutura semântica dos dados, a partir principalmente de medidas de redes complexas. Assim, este trabalho se diferencia dos citados, por contemplar um estudo para caracterizar e avaliar medidas de redes complexas em classificação de dados. Apesar de ser um trabalho original, na literatura existem alguns trabalhos que também buscam caracterizar redes complexas, embora não no contexto específico de classificação de dados como este trabalho, tais como:

- Em Newman (2003) analisa-se trabalhos sobre a estrutura e função de sistemas complexos (redes sociais, redes de informação, redes tecnológicas e redes biológicas). Com várias propriedades e medidas de redes complexas analisadas, o trabalho destaca medidas com boa aceitação pela comunidade.
- Em Rodrigues (2007), apresenta uma gama de informações sobre a teoria de redes complexas e principalmente das medidas de redes complexas, assim, ele caracteriza, classifica e analisa as redes complexas, sobre algumas bases específicas e reais, utilizando de métodos de estatística multivariada e reconhecimento de padrões.
- Relacionado com o trabalho anterior, em Costa et al. (2007) utiliza-se de medidas de redes complexas para apresentação e análise de um grande conjunto de medidas de redes complexas.

Materiais e Métodos

Este capítulo descreve as principais etapas para o desenvolvimento deste trabalho com o principal objetivo de caracterizar as medidas de redes complexas, descritas na seção 2.2, para a classificação de dados. Assim, a Seção 3.1 apresenta como foi realizada a geração das bases de dados artificiais e descreve as propriedades das bases reais selecionadas, a Seção 3.2 defini o algoritmo para a avaliação das medidas de redes complexas e a Seção 3.3 apresenta detalhes sobre o ambiente experimental definido para a realização dos experimentos.

3.1 Bases de dados

Para o estudo das medidas de redes complexas foi feita uma análise em dois tipos de bases: artificiais e reais. As bases artificiais foram geradas considerando diferentes níveis de ruídos nos dados, de modo a permitir uma análise mais coerente das medidas. Vale ressaltar que ruídos são registros inconsistentes na base de dados.

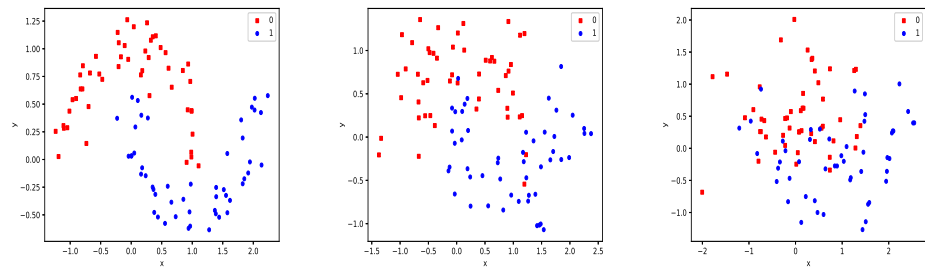
3.1.1 Bases artificiais

Foram geradas bases artificiais considerando os três domínios listados abaixo. Para todas elas, foram considerados três níveis de ruído, denotados como baixo, médio e alto. Para a geração das bases artificiais foi utilizado o pacote scikit, na linguagem Python.

□ Domínio Moons

Este domínio representa dois semicírculos intercalados. É um conjunto de dados simples para visualizar algoritmos de agrupamento e classificação. As bases geradas a partir do domínio Moons, cada uma com um nível diferente de ruído (baixo, médio e alto), são mostradas na Figura 7.

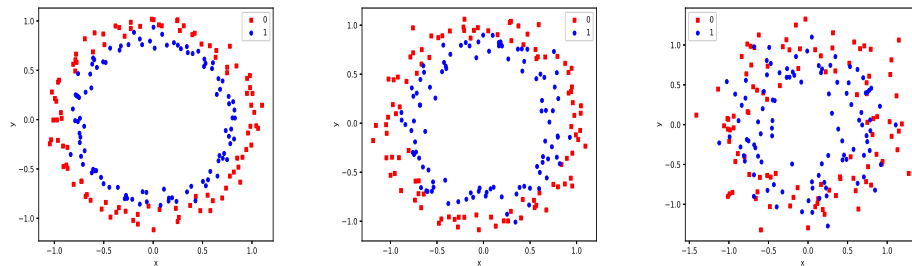
□ Domínio Circles



(a) Moons_b (nível de ruído baixo) (b) Moons_m (nível de ruído médio) (c) Moons_a (nível de ruído alto)

Figura 7 – Bases artificiais geradas a partir do domínio Moons, considerando três níveis de ruídos, denotados como baixo, médio e alto

O domínio Circles também é um conjunto de dados simples para visualizar algoritmos de agrupamento e classificação. Ele consiste basicamente de dois círculos concêntricos, ou seja, um "dentro" do outro. As bases geradas a partir do domínio Circles, são mostradas na Figura 8, cada uma com um nível diferente de ruído (baixo, médio e alto).



(a) Circles_b (nível de ruído baixo) (b) Circles_m (nível de ruído médio) (c) Circles_a (nível de ruído alto)

Figura 8 – Bases artificiais geradas a partir do domínio Circle, considerando três níveis de ruídos, denotados como baixo, médio e alto

□ Domínio Classification

O domínio Classification representa conjuntos de dados com classes ou grupos de objetos seguindo distribuição Gaussiana. As bases geradas a partir deste domínio, cada uma com um nível diferente de ruído (baixo, médio e alto), são mostradas na Figura 9.

3.1.2 Bases reais

Foram selecionadas quatro bases de dados reais para os experimentos, a saber: Appendicitis, Balance, Iris e Sonar, todas elas disponíveis no UCI Machine Learning Repository

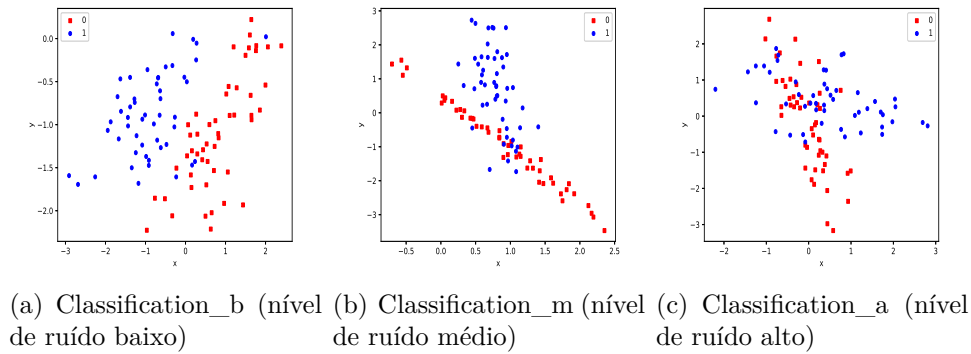


Figura 9 – Bases artificiais geradas a partir do domínio Classification, considerando três níveis de ruídos, denotados como baixo, médio e alto

(DUA; GRAFF, 2017). As características de cada base de dados: quantidade de instâncias (Instâncias), quantidade atributos (Atributos) e quantidade de classes (Classes) são exibidas na Tabela 1. A seguir é apresentada uma breve introdução sobre cada base de dados.

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

Nome	Instâncias	Atributos	Classes
Appendicitis	106	7	2
Balance	625	4	3
Iris	150	4	3
Sonar	208	4	3

□ Appendicitis

Neste conjunto de dados, os atributos representam sete medidas médicas tomadas em 106 pacientes, onde a classificação diz se o paciente tem apendicite (rótulo de classe 1) ou não (rótulo de classe 0).

□ Balance

A base Balance foi gerada para modelar experimentos psicológicos. O conjunto de dados possui 625 instâncias, com quatro atributos e contém três classes. Cada objeto é classificado como tendo a ponta da balança à direita, a ponta à esquerda ou equilibrada.

□ Iris

A base Iris contém 3 classes de 50 instâncias cada, portanto, 150 instâncias. Cada objeto possui quatro atributos, o comprimento da sépala em cm, a largura da sépala

em cm, o comprimento da pétala em cm e, a largura da pétala em cm, e classificadas como Íris Setosa, Íris Versicolour e Íris Virginica.

❑ Sonar

Na base Sonar, cada objeto da base é classificado com base nos atributos, como rocha ou cilindro de metal. A base contém 208 instâncias e cada objeto é um conjunto de 60 atributos.

3.2 Algoritmo para avaliação das medidas de rede

Para que seja possível a análise das medidas de redes complexas, primeiramente é necessário a construção da rede a partir dos dados na forma de vetor de atributos. Sobre tal grafo, a caracterização das medidas de rede será realizada tomando como referência a classificação via conformidade padrão.

A construção da rede a partir de dados do tipo vetor de atributos, faz-se utilizando o método de formação de rede K-vizinhos mais próximos (KNN). No trabalho, foi empregada a distância euclidiana como medida de dissimilaridade.

Construído o grafo, as medidas de redes complexas são aplicadas isoladamente sobre o sub-grafo correspondente a cada classe, de modo a caracterizá-lo antes da inserção de qualquer objeto de teste. Na etapa de teste, um novo objeto é inserido na rede e as medidas são recalculadas. O objeto recebe o rótulo da classe que apontou menor variação da medida. A seguir são detalhados os passos principais do algoritmo adaptado:

1. Divisão dos objetos em treino e teste utilizando validação cruzada;
2. Construção do grafo KNN usando os objetos de treino;
3. Cálculo das medidas de rede sobre os componentes relacionados a cada classe de dados;
4. Ao receber um objeto de teste, o mesmo é conectado temporariamente seguindo o mesmo critério para a formação do grafo;
5. Calcula-se novamente as medidas de redes complexas para as classes nas quais o novo objeto foi inserido;
6. Classifica o objeto de teste na classe em que ele causou menor variação da medida de rede.

3.3 Ambiente experimental

A implementação do algoritmo para classificação de dados utilizando medidas de redes complexas foi realizada na linguagem Python, utilizando das várias funções e pacotes existentes na linguagem.

Para a condução dos experimentos, foi utilizada a validação cruzada estratificada em 10 *fold*. Nela, o conjunto de dados é dividido em 10 pastas, de modo que a cada execução, 9 pastas sejam utilizadas para treino e 1 para teste. Além disso, o experimento foi repetido 3 vezes, alterando as configurações de objetos nas pastas, totalizando 30 execuções.

O único parâmetro considerado nas análises diz respeito à construção da rede KNN. Nesse sentido, variou-se K de 1 à 15 a fim de também avaliar a influência da rede no desempenho geral das medidas.

Por fim, vale ressaltar que este estudo analisa seis medidas de redes no contexto da classificação via conformidade de padrão, a saber: assortatividade, coeficiente de agrupamento, grau médio, intermedialidade, menor caminho médio e proximidade. O desempenho das medidas é avaliado em termos de acurácia média sobre as bases de dados artificiais e reais apresentadas neste capítulo.

Resultados

Neste capítulo será apresentado os resultados obtidos e as análises sobre o desempenho de cada medida de rede nas bases artificiais e reais apresentadas anteriormente. Especialmente, são analisadas seis medidas de redes complexas - assortatividade (R), coeficiente de agrupamento (CC), grau médio (K), intermedialidade (B), menor caminho médio (MC) e proximidade (P), todas apresentadas na Seção 2.2.

4.1 Resultados em bases artificiais

□ Domínio Moons

A Tabela 2 apresenta os resultados das medidas de redes para as bases de dados Moons. Os resultados referem-se aos valor mais altos de acurácia encontrado através da variação do parâmetro de construção da rede (k). Como pode ser observado na tabela, para os níveis de ruídos baixo e médio considerados, os resultados das medidas (e do baseline) são os mesmos. Contudo, ao analisar o cenário com nível de ruído alto, é possível observar uma melhora considerável em relação ao baseline, especialmente se considerarmos a medida de menor caminho médio, que alcançou 79% de acurácia.

Além da análise do melhor resultado preditivo, também analisamos a sensibilidade das medidas frente à variação dos parâmetros. Essa análise é importante, pois medidas mais robustas à variação dos parâmetros possuem vantagem em problemas complexos, onde a seleção de bons parâmetros pode ser difícil. A Figura 10, apresenta a variação da acurácia das medidas de redes complexas em função da variação do k . É possível verificar que as medidas grau médio, intermedialidade e proximidade decaem bastante com o incremento de k , principalmente na base de dados com maior nível de ruído. Comparada a elas, a assortatividade possui uma queda pequena nas bases com níveis de ruído baixo e médio, embora a queda também seja drástica para a base de dados com maior nível de ruído. Por outro lado, as medidas

Tabela 2 – Melhores valores de acurácia e desvio padrão das medidas de redes complexas para o domínio Moons. Entre parênteses é apresentado o valor de K usado para construir a rede.

Medidas de rede	Moons		
	Baixo ruído (b)	Médio ruído (m)	Alto ruído (a)
Assortatividade (R)	99.0±3.0 (k=1)	93.7±7.5 (k=1)	75.3±15.0 (k=3)
Coef. de agrup. (CC)	99.0±3.0 (k=1)	93.7±7.5 (k=1)	74.0±17.0 (k=3)
Grau médio (K)	99.0±3.0 (k=1)	93.7±7.5 (k=1)	75.3±16.7 (k=3)
Intermedialidade (B)	99.0±3.0 (k=1)	93.7±7.5 (k=1)	77.0±14.2 (k=5)
Men. cam. médio (MC)	99.0±3.0 (k=1)	93.7±7.5 (k=1)	79.0±13.0 (k=11)
Proximidade (P)	99.0±3.0 (k=1)	93.7±7.5 (k=1)	73.0±15.3 (k=3)
Baseline (1NN)	99.0±3.0	93.7±7.5	70.0 ± 15.7

do coeficiente de agrupamento e menor caminho médio demonstram ser mais robustas quanto à variação do k neste domínio, especialmente o menor caminho médio que consegue inclusive melhorar o seu desempenho preditivo na base de dados com nível de ruído maior.

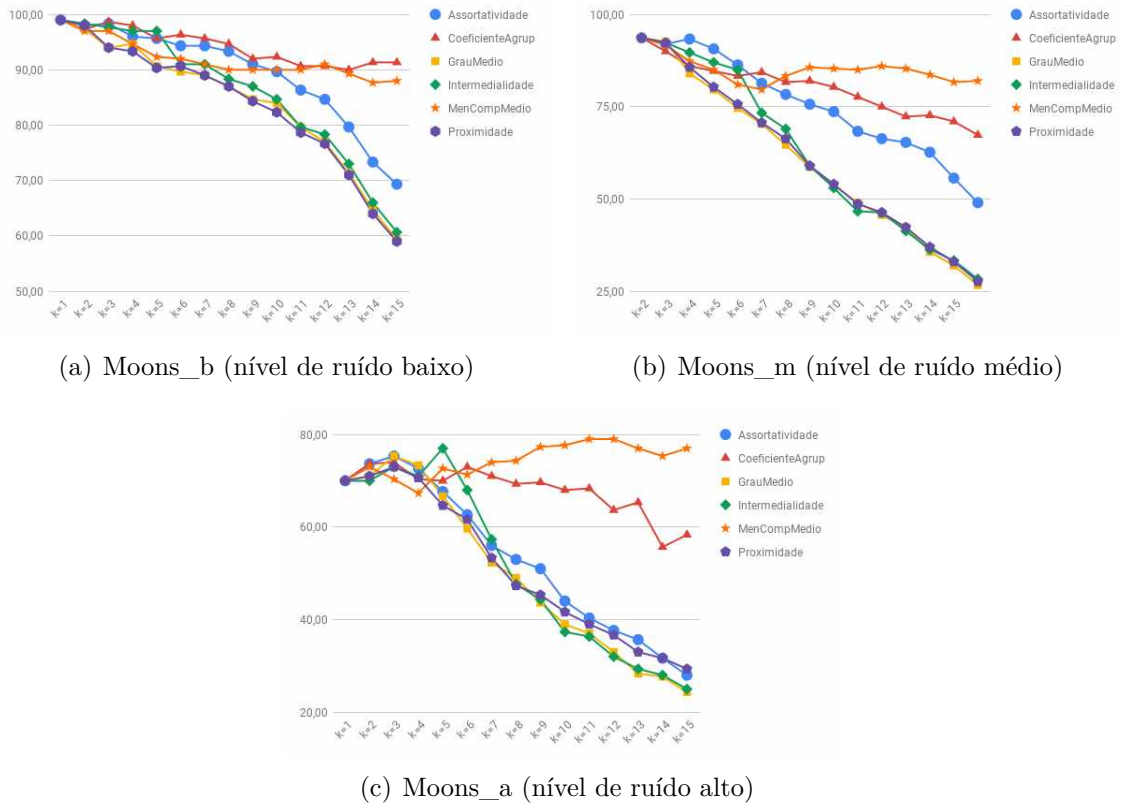


Figura 10 – Variação da acurácia das medidas de redes complexas para as redes do domínio Moons, e cada uma indicando o nível de ruído (baixo, médio e alto), em função de k.

Como forma de sumarmos a análise da influência do parâmetro k, a Tabela 3

apresenta a média do desempenho preditivo das medidas de rede considerando os 15 valores de k assumidos na construção do grafo. Nesse cenário, fica ainda mais evidente como as duas medidas discutidas anteriormente, coeficiente de agrupamento e menor caminho médio, se destacam frente às demais.

Na rede com baixo nível ruído a melhor medida é o Coeficiente de agrupamento na acurácia com 94.2%, e com 3.2% no desvio padrão. Na rede com médio nível ruído a melhor medida é o Menor Caminho Médio com 85.1% de acurácia e com 3.7% no desvio padrão. Na rede com alto nível ruído a medida Menor Caminho Médio tem o melhor resultado na acurácia com 74.4%, e com 3.5% no desvio padrão.

Tabela 3 – Acurácia e desvio padrão das medidas de redes complexas para o domínio Moons, considerando todas as variações do parâmetro de formação da rede (k).

Medidas de rede	Moons		
	Baixo ruído (b)	Médio ruído (m)	Alto ruído (a)
Assortatividade (R)	89.5±9.2	75.5±14.2	53.3±16.4
Coef. de agrup. (CC)	94.2±3.2	80.1±7.4	68.0±5.3
Grau médio (K)	84.1±11.8	60.2±21.8	50.0±18.2
Intermedialidade (B)	85.9±12.2	62.3±23.3	51.1±19.1
Men. cam. médio (MC)	91.9±3.4	85.1±3.7	74.4±3.5
Proximidade (P)	83.8±12.0	60.8±21.7	51.2±16.0

□ Domínio Circles

Na Tabela 4 tem-se os melhores resultados obtidos pelas medidas de rede para as bases de dados do domínio Circles. Assim como observado no domínio Moons, os resultados para os níveis de ruído baixo e médio são em sua maioria iguais ao do baseline, com excessão do Grau médio na base com nível de ruído baixo. Na rede com nível de ruído alto, todas as medidas foram capazes de melhorar os resultados do baseline, especialmente a proximidade com 61.2% de acurácia.

A Figura 11 apresenta a variação da acurácia das medidas de redes complexas tomando em consideração a variação do parâmetro k . Logo, podemos observar a sensibilidade das medidas com essa variação. Considerando a figura, pode-se notar que as medidas assortatividade, grau médio, intermedialidade e proximidade sofrem bastante, decaindo o desempenho preditivo a medida em que k é incrementado. Na base com maior nível de ruído, a proximidade tem seu melhor desempenho com o valor de $k = 2$, porém não consegue manter esse desempenho com o aumento de k , pelo contrário, para $k = 4$ ela apresenta a pior acurácia média dentre todas as medidas consideradas. O coeficiente de agrupamento, por sua vez, é menos sensível a variação de k , inclusive sendo capaz de melhorar seu desempenho preditivo com o incremento de k na base de dados com maior nível de ruído.

Tabela 4 – Melhores valores de acurácia e desvio padrão das medidas de redes complexas para o domínio Circles. Entre parênteses é apresentado o valor de K usado para construir a rede.

Medidas de rede	Circles		
	Baixo ruído (b)	Médio ruído (m)	Alto ruído (a)
Assortatividade (R)	89.3±6.4 (k=1)	82.7±6.5 (k=1)	56.5±10.5 (k=2)
Coef. de agrup. (CC)	89.3±6.4 (k=1)	82.7±6.5 (k=1)	57.3±9.0 (k=6)
Grau médio (K)	90.2±5.2 (k=2)	82.7±6.5 (k=1)	56.5±10.5 (k=2)
Intermedialidade (B)	89.3±6.4 (k=1)	82.7±6.5 (k=1)	56.5±10.5 (k=2)
Men. cam. médio (MC)	89.3±6.4 (k=1)	82.7±6.5 (k=1)	57.5±8.6 (k=12)
Proximidade (P)	89.3±6.4 (k=1)	82.7±6.5 (k=1)	61.2±9.5 (k=2)
baseline (1NN)	89.3±6.4	82.7±6.5	55.3±9.6

De modo geral, a medida menor caminho médio foi a mais consistente para este domínio, apresentando o melhor desempenho preditivo para as bases com nível de ruído baixo e médio, além de obter desempenho competitivo na base de dados com maior nível de ruído. Aliás, nesta base é possível verificar que a medida apresentou um pouco mais de sensibilidade em relação à variação do k, diferente das bases anteriores.

Assim como fizemos para o domínio Moons, também analisamos o desempenho das medidas de redes complexas considerando a média de todos os valores assumidos pelo parâmetro k, apresentado na Tabela 5. É possível observar que a medida menor caminho médio tem resultados muito melhores que as outras medidas para as bases com níveis de ruído baixo e médio. Considerando a base com maior nível de ruído, a medida com melhor resultado é o coeficiente de agrupamento, seguido de perto pelo próprio menor caminho médio.

Tabela 5 – Acurácia e desvio padrão das medidas de redes complexas para o domínio Circles, considerando todas as variações do parâmetro de formação da rede (k).

Medidas de rede	Circles		
	Baixo ruído (b)	Médio ruído (m)	Alto ruído (a)
Assortatividade (R)	64.8±13.9	58.3±13.2	46.7±5.9
Coef. de agrup. (CC)	52.5±20.8	54.4±13.8	54.3±2.3
Grau médio (K)	42.1±25.9	41.3±24.7	44.2±5.5
Intermedialidade (B)	37.5±31.0	62.3±23.3	45.1±7.3
Men. cam. médio (MC)	74.5±8.4	68.2±6.0	53.2±3.2
Proximidade (P)	46.3±23.3	46.4±17.7	48.8±4.4

□ Domínio Classification

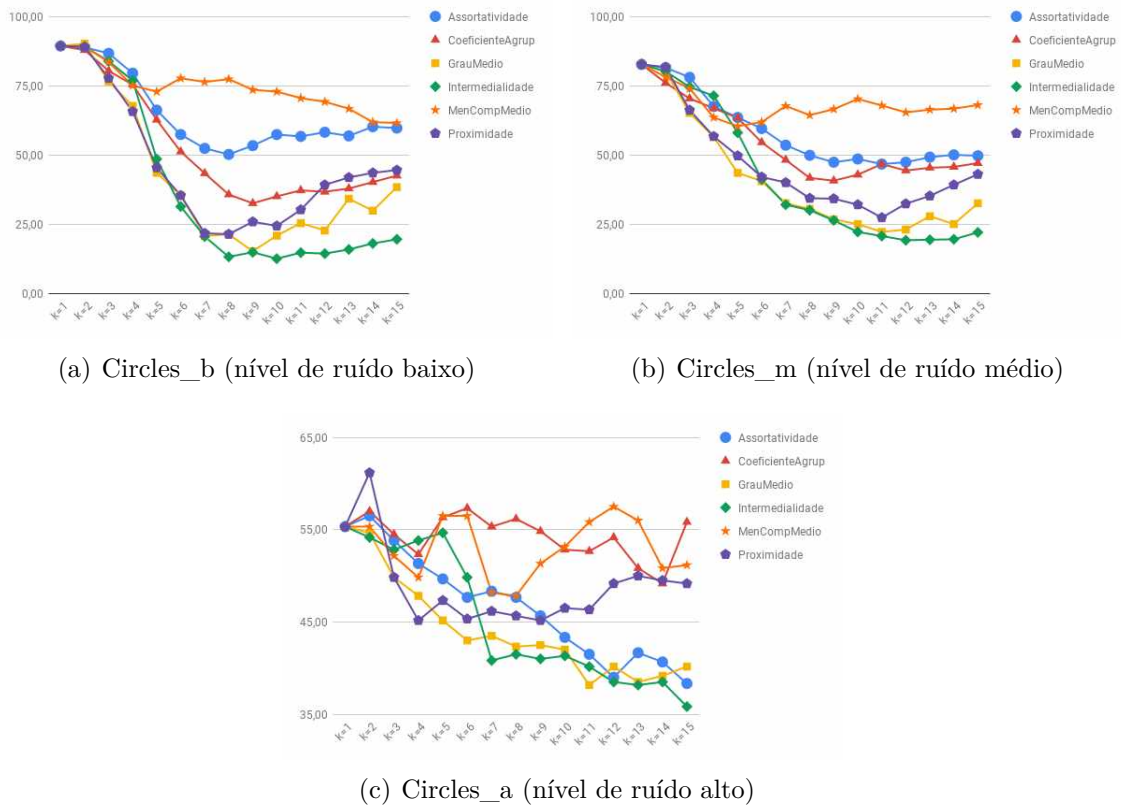


Figura 11 – Variação da acurácia das medidas de redes complexas para as redes do domínio Circles, e cada uma indicando o nível de ruído (baixo, médio e alto), e mostra qual medida de redes complexas tem melhor resultado para as bases.

A Tabela 6 apresenta os resultados de acurácia sobre o parâmetro de formação da rede k , para as bases do domínio Classification. Diferente dos domínios anteriores, aqui pode-se observar que nas três bases consideradas, o desempenho preditivo do baseline foi melhorado, embora apenas na base com maior nível de ruído isso tenha sido alcançado por todas as medidas. Na base com nível baixo de ruído, o grau médio alcançou 94.7% de acurácia. Na base com nível médio de ruído, a medida que superou o baseline é o menor caminho médio com 87.3% de acurácia. Na base com maior nível de ruído, o coeficiente de agrupamento atingiu o melhor desempenho, com 78.0% de acurácia.

Na Figura 12, observa-se a sensibilidade das medidas de redes complexas quanto à variação do parâmetro de construção da rede (K). Uma análise simples nos permite afirmar que as medidas assortatividade, grau médio, intermedialidade, e proximidade decaem muito mais com a variação de k do que o menor caminho médio e o coeficiente de agrupamento. Ambas demonstram uma certa linearidade nos resultados para base com menor nível de ruído. Por outro lado, ambas também apresentam uma melhora relevante de desempenho na base com maior nível de ruído, principalmente o coeficiente de agrupamento.

Tabela 6 – Melhores valores de acurácia e desvio padrão das medidas de redes complexas para o domínio Classification. Entre parênteses é apresentado o valor de K usado para construir a rede.

Medidas de rede	Classification		
	Baixo ruído (b)	Médio ruído (m)	Alto ruído (a)
Assortatividade (R)	92.7±6.3 (k=1)	86.0±1.1 (k=1)	71.7±1.2 (k=2)
Coef. de agrup. (CC)	92.7±6.3 (k=1)	86.0±1.1 (k=1)	78.0±1.2 (k=8)
Grau médio (K)	94.7±6.7 (k=2)	86.0±1.1 (k=1)	71.7±1.3 (k=2)
Intermedialidade (B)	94.3±7.2 (k=2)	86.0±1.1 (k=1)	72.7±1.5 (k=2)
Men. cam. médio (MC)	92.7±6.3 (k=1)	87.3±1.1 (k=2)	74.7±1.5 (k=7)
Proximidade (P)	94.0±7.6 (k=2)	86.0±1.1 (k=1)	71.3±1.4 (k=2)
baseline (1NN)	92.7±6.3	86.0±11.1	68.0±14.2

A Tabela 7 apresenta o desempenho preditivo das medidas de rede considerando a média dos resultados obtidos por todos os valores assumidos por k. Comparado com outras medidas, o coeficiente de agrupamento obteve os melhores resultados nas três bases de dados.

Tabela 7 – Acurácia e desvio padrão das medidas de redes complexas para o domínio Classification, considerando todas as variações do parâmetro de formação da rede (k).

Medidas de rede	Classification		
	Baixo ruído (b)	Médio ruído (m)	Alto ruído (a)
Assortatividade (R)	75.9±12.5	71.1±10.8	55.7±10.0
Coef. de agrup. (CC)	83.4±4.0	78.0±4.7	71.8±4.5
Grau médio (K)	70.4±13.7	66.2±12.2	51.9±10.1
Intermedialidade (B)	72.9±15.1	66.8±13.3	56.4±10.6
Men. cam. médio (MC)	82.5±4.2	77.1±4.8	66.8±4.3
Proximidade (P)	71.0±13.0	68.2±10.8	56.8±9.5

4.2 Resultados em bases reais

A Tabela 8, apresenta os resultados obtidos pelas medidas de redes para as bases reais Appendicitis, Balance, Iris e Sonar. Os resultados referem-se ao melhor valor de acurácia encontrado através da variação do parâmetro de construção da rede (k). Como pode ser observado nas quatro bases analisadas, as medidas superam o baseline. Na base Appendicitis a medida menor caminho médio tem o maior valor com 83.4% de acurácia. Na base Balance a medida grau médio com 95.4% de acurácia. Na base Iris a medida assortatividade é a melhor com 98.2% de acurácia. Na base Sonar a melhor medida com 84.7% de acurácia é o grau médio.

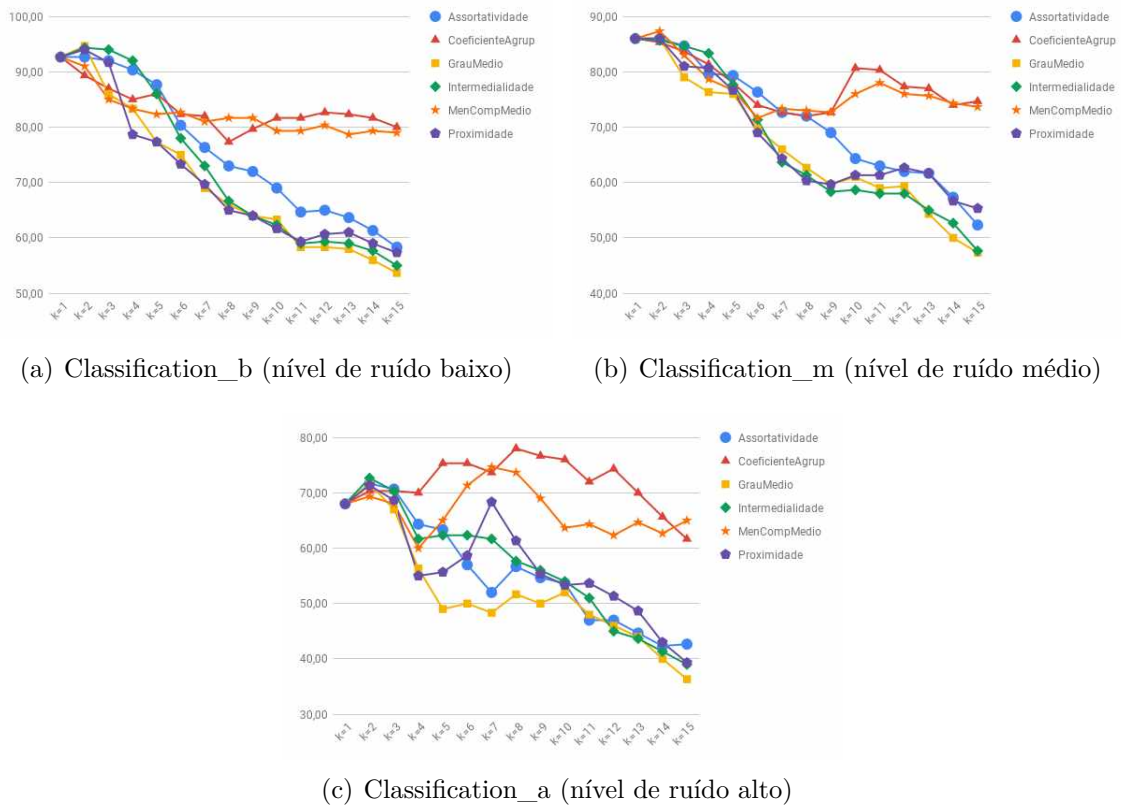


Figura 12 – Variação da acurácia das medidas de redes complexas para as redes do domínio Classification, e cada uma indicando o nível de ruído (baixo, médio e alto), e mostra qual medida de redes complexas tem melhor resultado para as bases.

Tabela 8 – Melhores valores de acurácia e desvio padrão das medidas de redes complexas para para as bases reais (Appendicitis, Balance, Iris e Sonar). Entre parênteses é apresentado o valor de K usado para construir a rede.

Medida de rede	Appendicitis	Balance	Iris	Sonar
Assortatividade (R)	82.8±6.2 (k=2)	95.3±2.5 (k=2)	98.2±3.4 (k=2)	83.5±9.6 (k=1)
Coef. de agrup. (CC)	80.8±8.1 (k=1)	94.8±2.8 (k=1)	96.4±4.8 (k=1)	83.5±9.6 (k=1)
Grau médio (K)	82.8±6.2 (k=2)	95.4±2.3 (k=2)	97.6±4.0 (k=2)	84.7±7.9 (k=2)
Intermedialidade (B)	80.8±8.1 (k=1)	94.8±2.8 (k=1)	97.6±4.0 (k=2)	83.5±9.6 (k=1)
Men. cam. médio (MC')	83.4±6.9 (k=9)	94.8±2.8 (k=1)	97.3±3.3 (k=15)	83.5±9.6 (k=1)
Proximidade (P)	83.1±6.3 (k=2)	95.0±2.5 (k=2)	97.6±4.0 (k=2)	83.5±9.6 (k=1)
Basel.(1NN)	80.8±8.1	94.8±2.8	96.4±4.8	83.5±9.6

Analizamos também a sensibilidade das medidas frente à variação dos parâmetros. A Figura 13, apresenta a variação da acurácia das medidas de redes complexas em função da variação do k, para as bases reais Appendicitis, Balance, Iris e Sonar. Na base Appendicitis (Figura 13(a)), as medidas que se destacam são assortatividade, coeficiente de agrupamento e, principalmente, menor caminho médio, enquanto a proximidade, o grau médio e a intermedialidade decaem a medida que k é incrementado. Na base Balance (Figura 13(b)), o menor caminho médio se mantém com

pouca variação, enquanto o coeficiente de agrupamento e assortatividade registram uma queda um pouco mais significativa, porém ainda distante da considerável queda de desempenho das medidas grau médio, proximidade e intermedialidade. Na base Iris (Figura 13(c)), o menor caminho médio tem o melhor resultado, visto que seu desempenho preditivo melhora com valores maiores de k . Mais uma vez, as medidas grau médio, intermedialidade e proximidade decaem bastante quando o valor de k aumenta, enquanto coeficiente de agrupamento e assortatividade apresentaram desempenho intermediário. Na base Sonar (Figura 13(d)), o menor caminho médio se mostra mais uma vez a medida mais robusta à variação de k , seguido pelo coeficiente de agrupamento e assortatividade. Novamente, as medidas grau médio, intermedialidade e proximidade se mostraram mais sensíveis à variação do parâmetro.

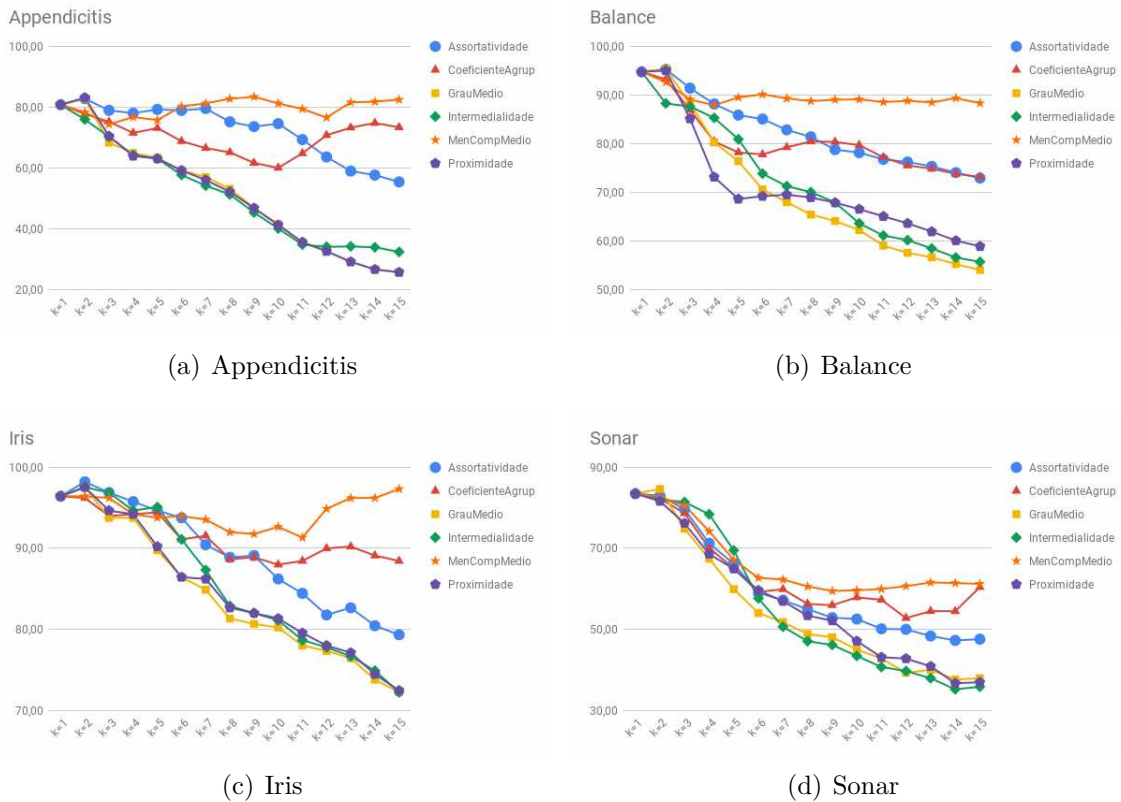


Figura 13 – Variação da acurácia das medidas de redes complexas para as bases reais (Appendicitis, Balance, Iris e Sonar).

A Tabela 9 apresenta a média do desempenho preditivo das medidas de rede considerando todos os valores de k assumidos na construção do grafo. Na tabela, é possível observar que o menor caminho médio apresenta melhor desempenho preditivo médio para as quatro bases, seguido pelo coeficiente de agrupamento e assortatividade.

Tabela 9 – Acurácia e desvio padrão das medidas de redes complexas para as bases reais (Appendicitis, Balance, Iris e Sonar), considerando todas as variações do parâmetro de formação da rede (k).

Medida de rede	Appendicitis	Balance	Iris	Sonar
Assortatividade (R)	72.4±9.2	82.5±7.4	89.3±6.5	60.2±13.2
Coef. de agrup. (CC)	70.5±5.9	80.4±6.5	91.3±3.0	63.2±10.4
Grau médio (K)	51.1±19.0	69.8±14.1	84.2±8.4	54.4±16.2
Intermedialidade (B)	51.4±16.7	71.7±12.9	85.7±8.9	55.3±18.5
Men. cam. médio (MC)	79.8±2.8	89.6±1.8	94.5±2.0	66.5±9.0
Proximidade (P)	51.0±19.1	71.2±11.4	84.9±8.2	56.3±15.8

4.3 Discussão dos resultados

Tanto nas bases artificiais quanto reais, dois tipos de análises são conduzidas: melhor desempenho preditivo e sensibilidade em relação à variação do parâmetro k . Na primeira análise, os resultados evidenciam, tanto para bases artificiais quanto reais, a dificuldade em avaliar as medidas de considerando apenas o melhor desempenho preditivo, pois as diferentes características das medidas poderão ser mais adequadas para um ou outro problema. Por exemplo, assortatividade alcançou melhor desempenho para a base de dados Iris, coeficiente de agrupamento para a base Classification (alto), grau médio para as bases Circles (baixo), Classification (baixo), Balance e Sonar, menor caminho médio para as bases Moons (alto), Classification (médio) e Appendicitis, e proximidade para a base Circles (alto). Logo, essa grande variabilidade em termos de desempenho não abre espaço para muitas conclusões à respeito dessa análise.

Na segunda análise, relacionada à sensibilidade das medidas em relação à variação de k , os resultados são mais conclusivos. Medidas como coeficiente de agrupamento e menor caminho médio apresentaram menor sensibilidade à escolha do parâmetro. Por exemplo, menor caminho médio alcançou melhor desempenho para as bases de dados Moons (médio e alto), Circles (baixo e médio), Appendicitis, Balance, Iris e Sonar, coeficiente de agrupamento para as bases Moons (baixo), Circles (alto) e Classification (baixo, médio e alto). Essa é uma análise importante, pois medidas mais robustas à variação dos parâmetros possuem vantagem em problemas complexos, onde a seleção de bons parâmetros pode ser difícil.

A seguir apresentamos algumas discussões sobre as medidas consideradas no estudo:

- A assortatividade não obteve os melhores resultados, porém conseguiu se manter em nível intermediário diante dos cenários analisados. A assortatividade verifica se ocorre a ligação entre nós de grau similar ou não, o que pode ser uma medida importante caso as classes apresentem padrões de assortatividade distintos.
- A medida coeficiente de agrupamento obteve bons resultados, principalmente no

domínio Classification. Essa medida pode ser interessante em bases de dados cujas classes apresentem diferentes tendências em relação à formação de grupos.

- ❑ O grau médio não obteve bons resultados para a análise de sensibilidade, mas quando analisamos a medida em termos de desempenho preditivo, ela obteve o melhor resultado em algumas redes, especialmente aquelas com valores de k menores. Apesar da simplicidade da medida, acredita-se que ela pode ser especialmente útil em bases de dados fortemente definidas por características locais dos dados, bem representadas pelo número de conexões.
- ❑ A medida de intermedialidade não obteve bons resultados em ambas as análises. Ademais, possui o maior custo computacional dentre todas as medidas. Dessa forma, considerando os tipos de dados investigados, não recomendamos o uso dessa medida.
- ❑ O menor caminho médio foi a medida com melhores resultados considerando ambas as análises. A medida calcula a média de menor caminho de cada nó da rede para todos os outros, o que parece ser uma estratégia interessante para caracterizar a conformidade de padrão após a inserção de um novo objeto e também para ser menos sensível à variação do parâmetro relacionado ao número de conexões.
- ❑ A proximidade mede o quão próximo os nós de uma rede estão de todos os outros. A medida obteve desempenho preditivo razoável se comparada com as demais, embora não tenha apresentado bons resultados na análise de sensibilidade. De fato, nossos resultados indicam que, no contexto da conformidade de padrão, considerar a proximidade média da rede parece ser uma abordagem menos apropriada do que o menor caminho médio.

Conclusão

Este trabalho teve como objetivo principal avaliar comparativamente um conjunto de medidas de redes complexas para a classificação de dados. Apesar de se tornarem cada vez mais populares neste contexto, a literatura ainda não conta com um estudo voltado para análise das características das medidas frente a diferentes domínios de dados. A partir de uma análise da literatura, foram selecionadas medidas de redes complexas relevantes para o contexto da classificação de dados, a saber: assortatividade, coeficiente de agrupamento, grau médio, intermedialidade, menor caminho médio e proximidade. Em seguida, para projetar e desenvolver um ambiente experimental propício para caracterizar empiricamente as medidas de redes tanto em cenários artificiais com diferentes níveis de ruído quanto em bases reais de diferentes domínios, necessitou-se adaptar o algoritmo de classificação via conformidade de padrão para permitir a análise comparativa das medidas de redes complexas individualmente.

No capítulo 4 foram apresentados os resultados dos experimentos com o objetivo de avaliar e caracterizar o comportamento das medidas de redes complexas para a classificação de dados. Diante disso, dois tipos de análise foram conduzidas: melhor desempenho preditivo e sensibilidade à variação do parâmetro de construção da rede (k).

Na primeira análise, foi observado que os melhores desempenhos foram alcançados por diferentes medidas, a depender das bases de dados considerada. Nesse sentido, o trabalho demonstra a dificuldade em avaliar as medidas considerando apenas o melhor desempenho preditivo. De fato, esse resultado evidencia que as diferentes características relacionadas às medidas selecionadas podem torná-las mais adequadas para um ou outro problema.

Por outro lado, a segunda análise onde avalia-se a sensibilidade das medidas em relação à variação na formação do grafo, mostra que as medidas como coeficiente de agrupamento e menor caminho médio apresentaram menor sensibilidade, isto é, essas medidas são mais robustas à variação do parâmetro. De certa forma, os resultados demonstram que medidas como grau médio, proximidade e intermedialidade são muito mais sensíveis à variação do parâmetro k .

Outras análises foram ainda apresentadas relacionadas às características específicas

das medidas sob estudo, o que espera-se que contribua com a literatura sobre o assunto.

Para trabalhos futuros pretende-se considerar um número maior de bases de dados artificiais e reais. Também pretende-se analisar mais medidas de redes complexas, visto que há uma quantidade significativa de medidas na literatura. Também pretende-se avaliar outros métodos de construção da rede, além do grafo KNN.

Referências

AMANCIO, D. R. **Classificação de textos com redes complexas**. 46 p. Tese (Doutorado) — Instituto de Física de São Carlos, São Carlos - SP, 2013. Citado na página 22.

ARAÚJO, B. M. de. **Rotulação de indivíduos representativos no aprendizado semissupervisionado baseado em redes: caracterização, realce, ganho e filosofia**. 24-26 p. Tese (Doutorado) — Instituto de Ciências Matemáticas e de Computação (ICMC-USP), São Carlos - SP, 2015. Citado na página 24.

BARBIERI, A. L. **Análise de Robustez em Redes Complexas**. Dissertação (Mestrado) — Instituto de Física de São Carlos, São Carlos - SP, 2010. Citado na página 21.

Bertini Jr., J. R. **Classificação de dados estacionários e não estacionários baseada em grafos**. Tese (Doutorado) — Instituto de Ciências Matemáticas e de Computação (ICMC-USP), São Carlos - SP, 2010. Citado na página 24.

CARNEIRO, M.; ZHAO, L. Analysis of graph construction methods in supervised data classification. In: IEEE. **Brazilian Conference on Intelligent Systems**. [S.l.], 2018. p. 390–395. Citado 4 vezes nas páginas 14, 22, 23 e 24.

CARNEIRO, M. G. **Redes complexas para classificação de dados via conformidade de padrão, caracterização de importância e otimização estrutural**. Tese (Doutorado) — Instituto de Ciências da Computação (ICMC/USP), São Carlos - SP, 2016. Citado 5 vezes nas páginas 13, 20, 21, 22 e 23.

CARNEIRO, M. G. et al. Particle swarm optimization for network-based data classification. **Neural Networks**, Elsevier, v. 110, p. 243–255, 2019. Citado 3 vezes nas páginas 13, 14 e 24.

_____. Nature-inspired graph optimization for dimensionality reduction. In: IEEE. **IEEE International Conference on Tools with Artificial Intelligence**. [S.l.], 2017. p. 1113–1119. Citado na página 24.

_____. Network-based data classification: combining k-associated optimal graphs and high-level prediction. **Journal of the Brazilian Computer Society**, Springer, v. 20, n. 1, p. 1–14, 2014. Citado na página 24.

- CARNEIRO, M. G.; ZHAO, L. Organizational data classification based on the importance concept of complex networks. **IEEE Transactions on Neural Networks and Learning Systems**, IEEE, v. 29, n. 8, p. 3361–3373, 2018. Citado 4 vezes nas páginas 13, 14, 15 e 24.
- CARNEIRO, M. G. et al. Network structural optimization based on swarm intelligence for highlevel classification. In: IEEE. **IEEE International Joint Conference on Neural Networks**. [S.l.], 2016. p. 3737–3744. Citado na página 24.
- COSTA, L. da F. et al. Caracterização de redes complexas: Um levantamento de medidas. **Avanços em Física**, v. 56, n. 1, p. 167–242, 2007. Citado 2 vezes nas páginas 20 e 25.
- CUPERTINO, T. H. et al. A scheme for high level data classification using random walk and network measures. **Expert Systems with Applications**, Springer, v. 92, p. 289–303, 2018. Citado 3 vezes nas páginas 13, 14 e 24.
- DUA, D.; GRAFF, C. **UCI Machine Learning Repository**. 2017. Online; accessed 16-Novembro-2019. Disponível em: <<http://archive.ics.uci.edu/ml>>. Citado na página 28.
- FERRERO, C. A. **Algoritmo KNN para previsão de dados temporais: funções de previsão e critérios de seleção de vizinhos próximos aplicados a variáveis ambientais de limnologia**. Dissertação (Mestrado) — Instituto de Ciências da Computação (ICMC/USP), São Carlos - SP, 2009. Citado 2 vezes nas páginas 8 e 19.
- LOPES, G. A. W. **Um Modelo de Rede Complexa para Análise de Informações Textuais**. Dissertação (Mestrado) — Centro Universitário da FEI, 2011. Citado na página 13.
- NEWMAN, M. E. J. The structure and function of complex networks. São Carlos - SP, 2003. Citado 4 vezes nas páginas 8, 14, 20 e 25.
- RODRIGUES, F. A. **Caracterização, classificação e análise de redes complexas**. Tese (Doutorado) — Instituto de Física de São Carlos (USP), São Carlos - SP, 2007. Citado na página 25.
- 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, 2012. Citado 5 vezes nas páginas 8, 13, 14, 23 e 24.
- TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introdução ao Datamining: mineração de dados**. Rio de Janeiro: [s.n.], 2009. Citado na página 17.
- VERA, A. M. **Propriedades de Redes Complexas de Telecomunicações**. Dissertação (Mestrado) — Escola de Engenharia de São Carlos, São Carlos - SP, 2011. Citado na página 22.
- ZACHARY, W. An information flow model for conflict and fission in small groups. **Journal of Anthropological Research**, v. 33, n. 4, 1977. Citado 3 vezes nas páginas 8, 13 e 15.