UNIVERSIDADE FEDERAL DE SÃO CARLOS
Rod. Washington Luís, km 235 SP-310, CEP: 13565-905, São Carlos - SP

Avaliação de Métodos de Construção de Grafos para Classificação no Aprendizado Semi-Supervisionado

São Carlos - SP
2025

Agradecimentos

Agradeço primeiramente à minha família, pelo apoio incondicional, amor e incentivo ao longo de toda a minha trajetória acadêmica.

À Universidade Federal de São Carlos, pelos anos de formação e pelas oportunidades de aprendizado e crescimento pessoal e profissional.

Ao meu orientador, Professor Alan Valejo, por toda a orientação, paciência e comprometimento durante o desenvolvimento deste trabalho. Suas sugestões, ensinamentos e disponibilidade foram fundamentais para a conclusão deste projeto.

Aos colegas do curso de Engenharia Física, pela convivência, trocas de conhecimento e apoio mútuo ao longo dessa caminhada.

Por fim, agradeço a todos que, de alguma forma, contribuíram direta ou indiretamente para a realização deste trabalho.

Resumo

O aprendizado semi-supervisionado tem ganhado destaque em tarefas de classificação onde apenas uma fração das amostras possui rótulos disponíveis. Nesse contexto, métodos baseados em grafos têm se mostrado eficazes por explorarem a estrutura de similaridade entre os dados. Este trabalho investiga o impacto de diferentes estratégias de construção de grafos no desempenho da propagação de rótulos, comparando quatro abordagens: kNN (baseline), RGCLI, KAOG e SNGC. Os métodos foram avaliados em cinco conjuntos de dados, com diferentes proporções de dados rotulados e configurações de parâmetros. Os resultados demonstram que, embora o método kNN — que não utiliza informação supervisionada — tenha apresentado desempenho competitivo na maioria dos cenários, o RGCLI se destacou entre os métodos supervisionados, combinando robustez com boa capacidade de generalização. Observou-se ainda que o desempenho dos métodos varia de acordo com a complexidade e estrutura dos conjuntos de dados, sendo o SNGC mais eficaz em bases de alta dimensionalidade, e o KAOG limitado por sua baixa conectividade. A análise estatística confirmou diferenças significativas entre os métodos para todas as métricas avaliadas. Os achados reforçam a importância de considerar tanto a natureza dos dados quanto a estratégia de construção do grafo na escolha de técnicas semi-supervisionadas.

Palavras-chave: Grafos; Construção de Grafos; Aprendizado de Máquina;Aprendizado Semi-Supervisionado; Propagação de Rótulos; RGCLI; SNGC; KAOG.

Abstract

Semi-supervised learning has gained relevance in classification tasks where only a fraction of the samples are labeled. In this context, graph-based methods have proven effective by leveraging the similarity structure among data points. This work investigates the impact of different graph construction strategies on the performance of label propagation, comparing four approaches: kNN (baseline), RGCLI, KAOG, and SNGC. The methods were evaluated on five datasets with varying proportions of labeled data and parameter settings. Results show that although the kNN method — which does not use supervision — achieved competitive performance in most scenarios, RGCLI stood out among the supervised methods, combining robustness with good generalization. It was also observed that the performance of the methods varies according to the complexity and structure of the datasets, with SNGC being more effective on high-dimensional data, and KAOG limited by its low connectivity. The statistical analysis confirmed significant differences between the methods across all evaluated metrics. These findings reinforce the importance of considering both data characteristics and graph construction strategies when choosing semi-supervised learning techniques.

Keywords: Graphs; Graph Construction; Machine Learning; Semi-Supervised Learning; Label Propagation; RGCLI; SNGC; KAOG.

Sumário

Lista de ilustrações

1 INTRODUÇÃO

1.1 Contextualização

Enquanto a internet se torna cada vez mais presente em todas as esferas da sociedade — da vida pessoal às tecnologias utilizadas por empresas e governos — a quantidade de dados gerados e armazenados cresce de forma exponencial. Esses dados podem estar disponíveis publicamente, como em redes sociais, ou restritos a sistemas privados corporativos e institucionais. Diante desse cenário, extrair significado dessa imensa massa de informação torna-se um desafio central para diversas áreas do conhecimento.

A tarefa de transformar dados em informação útil é melhor conduzida por meio de algoritmos de aprendizado de máquina, uma subárea da inteligência artificial que busca desenvolver sistemas capazes de aprender padrões e tomar decisões com mínima intervenção humana. Esses algoritmos adaptam seu comportamento com base em experiências anteriores e têm sido amplamente aplicados em tarefas como reconhecimento de imagens, diagnósticos médicos, sistemas de recomendação, análise de sentimentos, entre outras.

O aprendizado de máquina pode ser dividido, de forma geral, em três paradigmas principais: aprendizado supervisionado, não supervisionado e semi-supervisionado. No aprendizado supervisionado, os algoritmos são treinados com exemplos rotulados, ou seja, com entradas associadas a saídas conhecidas, com o objetivo de classificar novas instâncias. No aprendizado não supervisionado, os dados não possuem rótulos e o modelo busca identificar padrões ou agrupamentos naturais. Já o aprendizado semi-supervisionado representa uma abordagem intermediária, em que um pequeno conjunto de dados rotulados é combinado com um grande volume de dados não rotulados. Essa abordagem busca aproveitar o conhecimento parcial contido nos dados rotulados para orientar o processo de aprendizagem, superando as limitações impostas pelo alto custo da rotulagem manual.

Dentro do aprendizado semi-supervisionado, destaca-se a tarefa de **classificação**, cujo objetivo é prever corretamente os rótulos das instâncias não rotuladas. Essa tarefa é particularmente relevante em contextos onde a obtenção de rótulos confiáveis é onerosa ou limitada, como em aplicações biomédicas, ambientais ou de segurança. Entre as diversas técnicas empregadas para essa finalidade, os métodos baseados em grafos têm se mostrado promissores, especialmente quando combinados com algoritmos de propagação de rótulos.

1.2 Motivação

A crescente disponibilidade de dados não rotulados em diversos domínios — como redes sociais, sistemas médicos, ambientes industriais e sensores ambientais — contrasta com a escassez de dados rotulados, cuja obtenção geralmente envolve custos elevados, conhecimento especializado ou limitações práticas. Nesse contexto, o aprendizado semi-supervisionado

apresenta-se como uma abordagem promissora, ao permitir que modelos computacionais façam uso de uma pequena fração de exemplos rotulados juntamente com um grande volume de dados não rotulados, melhorando a capacidade de generalização em tarefas de predição.

Dentre as abordagens desse paradigma, destaca-se o *aprendizado semi-supervisionado baseado em grafos*, no qual os dados são representados por uma estrutura de rede. Nessa representação, as instâncias correspondem aos vértices do grafo e as arestas indicam a similaridade entre pares de instâncias. A partir dessa estrutura, algoritmos de propagação de rótulos — como o *Label Propagation* e o *Label Spreading* — exploram a conectividade do grafo para inferir os rótulos das instâncias não rotuladas, assumindo que vértices conectados devem pertencer à mesma classe.

A eficácia desses algoritmos depende diretamente da topologia do grafo sobre o qual são aplicados. Isso faz com que a etapa de **construção do grafo** seja um fator crítico no desempenho da tarefa de classificação. Diferentes estratégias de construção — como k-Nearest Neighbors (k-NN), ε-Vizinhança, Mutual k-NN (MkNN), Relaxed Minimum Spanning Tree (RMST), B-Matching e RGCLI — produzem grafos com propriedades estruturais distintas, que impactam a forma como os rótulos são propagados.

Por exemplo, o método RGCLI (Robust Graph Construction with Labeled Instances) considera explicitamente os rótulos disponíveis ao construir a rede, buscando promover maior coesão entre instâncias da mesma classe e reduzir conexões ambíguas. Por outro lado, métodos mais tradicionais como o k-NN são baseados apenas em distância e ignoram a informação supervisionada, o que pode levar à criação de conexões problemáticas em regiões onde há sobreposição entre classes.

Apesar dessa relevância, a literatura ainda carece de estudos que analisem de forma sistemática o impacto que a escolha do método de construção do grafo exerce sobre o desempenho dos algoritmos de classificação semi-supervisionada. Essa lacuna motiva a presente pesquisa, que busca compreender em profundidade como diferentes estratégias de construção influenciam os resultados obtidos por modelos de propagação de rótulos em contextos semi-supervisionados.

1.3 Problema de Pesquisa

Diante do exposto, este trabalho se propõe a investigar a seguinte questão central:

Qual é o impacto da escolha do método de construção de grafos no desempenho de algoritmos de propagação de rótulos no contexto de aprendizado semi-supervisionado?

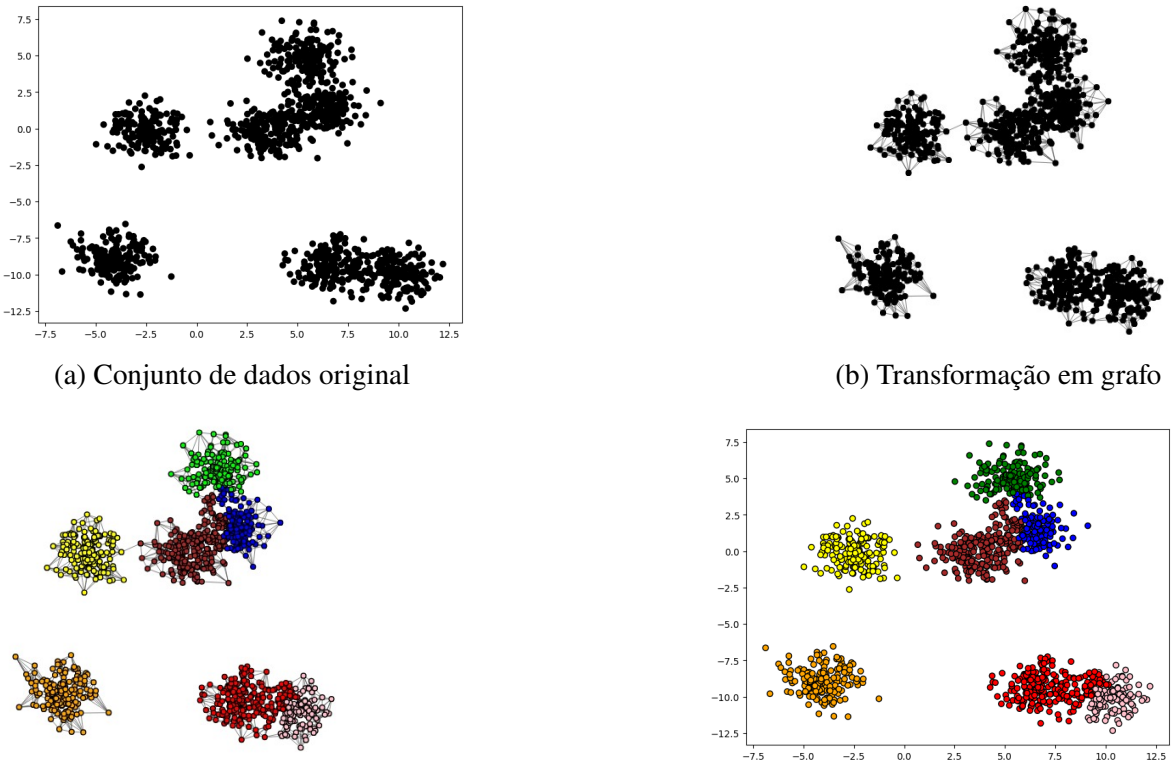
1.4 Objetivos

1.4.1 Objetivo Geral

O objetivo geral deste trabalho é realizar uma análise comparativa do desempenho de diferentes métodos de construção de grafos no contexto de aprendizado semi-supervisionado, avaliando como essas diferentes topologias influenciam os resultados de algoritmos de propagação de rótulos aplicados à tarefa de classificação.

No contexto deste trabalho, cada instância de dado é representada por um vértice, e as conexões entre instâncias são determinadas com base em alguma medida de similaridade, formando um grafo. A partir dessa estrutura, algoritmos de propagação de rótulos inferem os rótulos das instâncias não rotuladas a partir de poucas instâncias rotuladas, utilizando a conectividade do grafo como base para essa propagação. A Figura 1 ilustra esse processo, desde a representação inicial dos dados até a atribuição final dos rótulos inferidos.

Figura 1 – Processo de agrupamento com grafos.



(a) Conjunto de dados original

(b) Transformação em grafo

(c) Classificação

(d) Classes projetadas nos dados originais

Fonte: (Do Autor)

Considerando que diferentes métodos de construção de grafos podem gerar topologias significativamente distintas — influenciando, portanto, o desempenho da classificação — este trabalho busca investigar sistematicamente o impacto dessas escolhas sobre os resultados obtidos.

1.4.2 Objetivos Específicos

- Investigar e revisar a literatura científica sobre métodos de construção de grafos e algoritmos de aprendizado semi-supervisionado baseados em grafos;
- Implementar diferentes estratégias de construção de grafos e aplicar algoritmos de propagação de rótulos em conjuntos de dados sintéticos;
- Comparar o desempenho dos algoritmos *Label Propagation* e *Label Spreading* em diferentes grafos, utilizando métricas como acurácia, NMI e F1-score;
- Avaliar a influência das características topológicas dos grafos gerados (densidade, conectividade, modularidade) sobre o desempenho dos algoritmos;
- Identificar quais métodos de construção tendem a favorecer o desempenho da propagação de rótulos em diferentes cenários.

1.5 Justificativa

A crescente adoção de algoritmos baseados em grafos no aprendizado semi-supervisionado, aliada à diversidade de estratégias disponíveis para construir tais grafos, torna urgente a realização de estudos que explorem o impacto dessas escolhas no desempenho dos modelos. A maioria das pesquisas tende a focar nos algoritmos de propagação em si, assumindo a construção do grafo como uma etapa fixa ou secundária. No entanto, como este trabalho defende, a qualidade e as propriedades estruturais da rede construída são determinantes para o sucesso da tarefa de classificação semi-supervisionada.

Ao analisar comparativamente diferentes métodos de construção, este estudo busca preencher uma lacuna importante na literatura e fornecer subsídios práticos para pesquisadores e profissionais da área que desejam utilizar técnicas gráficas com maior efetividade. Além disso, o trabalho contribui para a formação acadêmica do discente, consolidando conhecimentos nas áreas de aprendizado de máquina, análise de grafos e experimentação científica.

1.6 Organização do Trabalho

Este trabalho está organizado da seguinte forma: Na Seção 1 é apresentada a proposta de pesquisa, os objetivos e a justificativa para realizá-lo. Na Seção 2 é realizada uma revisão bibliográfica, fazendo o levantamento de conceitos importantes de Aprendizagem de Máquina e os algoritmos utilizados nos experimentos. Na Seção 3 é apresentada a metodologia para realização dos experimentos, bem como as ferramentas utilizadas e bases de dados utilizados. A Seção 4 apresenta a análise dos resultados obtidos e a análise comparativa entre os algoritmos.

Por fim, na Seção 5 é descrito as conclusões do trabalho, bem como os próximos passos que podem ser explorados em trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Aprendizado de Máquina

O Aprendizado de Máquina (*Machine Learning*) é uma subárea da Inteligência Artificial que se concentra no desenvolvimento de algoritmos capazes de aprender automaticamente a partir de dados. Em vez de dependerem exclusivamente de regras pré-programadas, esses algoritmos constroem modelos capazes de extrair padrões e regularidades a partir de exemplos, com o objetivo de realizar previsões ou tomar decisões em novos contextos (MONARD et al., 1997).

O processo de aprendizado é majoritariamente indutivo: os algoritmos generalizam comportamentos ou estruturas com base em um conjunto finito de dados observados. Essa generalização é essencial tanto para prever novas situações quanto para compreender e organizar grandes volumes de dados complexos (REZENDE, 2003).

O aprendizado de máquina pode ser classificado em três paradigmas principais, de acordo com a disponibilidade de rótulos nos dados de entrada: aprendizado supervisionado, aprendizado não supervisionado e aprendizado semi-supervisionado.

2.1.1 Aprendizado Supervisionado

No aprendizado supervisionado, o modelo é treinado a partir de um conjunto de dados rotulados, isto é, para cada instância de entrada existe uma saída conhecida (classe ou valor). O objetivo do algoritmo é aprender uma função de mapeamento capaz de predizer corretamente a saída para novas entradas. Essa abordagem é amplamente utilizada em tarefas como classificação (atribuição de categorias) e regressão (previsão de valores contínuos) (KOTSIANTIS et al., 2007).

2.1.2 Aprendizado Não Supervisionado

O aprendizado não supervisionado, por sua vez, opera sobre dados não rotulados. Nessa abordagem, os algoritmos buscam identificar padrões, estruturas ou agrupamentos latentes nos dados, sem qualquer supervisão externa. Entre suas aplicações destacam-se o *clustering*, a detecção de anomalias e a redução de dimensionalidade, por meio de técnicas como a Análise de Componentes Principais (PCA) (GHAHRAMANI, 2003).

2.1.3 Aprendizado Semi-Supervisionado

O aprendizado semi-supervisionado (*Semi-Supervised Learning* – SSL) combina os dois paradigmas anteriores, operando sobre um conjunto de dados parcialmente rotulado. Essa abordagem é especialmente útil em cenários em que a rotulagem manual de dados é dispendiosa, demorada ou exige conhecimento especializado — como ocorre nas áreas da medicina, direito, segurança ou pesquisa científica (ZHU, 2005; FONTES, 2023).

Nesse cenário, busca-se utilizar a estrutura dos dados não rotulados como fonte de informação adicional para reforçar o aprendizado a partir dos poucos exemplos rotulados. Para isso, diversas estratégias têm sido propostas, incluindo:

- *Auto-training*: o modelo rotula iterativamente exemplos não rotulados com maior confiança e os reaproveita no treinamento (ZHOU; LI, 2005; IBM, 2023);
- *Co-training*: dois classificadores independentes rotulam os exemplos um do outro, assumindo que usam visões distintas dos dados (BLUM; MITCHELL, 1998; CALDAS, 2017);
- Técnicas baseadas em regularização e otimização (ZHU, 2005);
- Redes neurais com pseudo-rótulos gerados automaticamente (IBM, 2023);
- **Métodos baseados em grafos** (FONTES, 2023; SANTOS, 2014).

2.2 Aprendizado de Máquina com Grafos

A utilização de grafos no aprendizado de máquina tem se mostrado eficaz especialmente em cenários nos quais as relações entre os dados desempenham papel fundamental na inferência. Ao invés de considerar apenas instâncias isoladas, os métodos baseados em grafos incorporam a conectividade entre os dados, possibilitando a exploração de padrões locais e globais.

No contexto semi-supervisionado, essa representação gráfica permite que algoritmos de propagação de rótulos explorem a estrutura da rede para difundir os rótulos disponíveis, aumentando a abrangência da supervisão. A qualidade da propagação, novamente, está diretamente associada à maneira como o grafo é construído — justificando o foco deste trabalho na análise comparativa de métodos de construção.

2.2.1 Métodos Baseados em Grafos

Uma das abordagens mais promissoras para o aprendizado semi-supervisionado é aquela baseada em grafos. Nesses métodos, os dados são representados como uma rede, cujos vértices correspondem às instâncias e as arestas indicam relações de similaridade. Essa estrutura permite a aplicação de algoritmos de propagação de rótulos, como *Label Propagation* e *Label*

Spreading, que inferem os rótulos das instâncias não rotuladas explorando a conectividade do grafo. A eficácia desses algoritmos depende fortemente da topologia da rede, o que torna a **construção do grafo** uma etapa crítica no processo de aprendizado.

A seguir, são apresentados diversos métodos de construção de grafos propostos na literatura, com destaque para aqueles que utilizam informações supervisionadas na própria construção da rede:

- **SNGC (Supervised Neighborhood Graph Construction)** — Rohban & Rabiee (2012): constrói grafos penalizando ligações entre classes diferentes e otimizando a vizinhança com base em regularização e entropia cruzada, demonstrando bons resultados mesmo com poucos rótulos (ROHBAN; RABIEE, 2012).
- **Subgrafos homogêneos com seleção de atributos** — Berton & Lopes (2014): constrói subgrafos coesos para instâncias rotuladas e conecta as não rotuladas com base em densidade e similaridade (BERTON; LOPES, 2014).
- **Construção incremental e robusta** — Berton & Lopes (2015): adiciona aprendizado incremental, análise de robustez topológica e remoção de conexões ruidosas (BERTON; LOPES, 2015).
- **RGCLI (Robust Graph that Considers Labeled Instances)** — Berton & Lopes (2014): prioriza conexões entre instâncias da mesma classe rotulada e conecta não rotuladas com base em densidade e estrutura local (BERTON et al., 2017a).
- **LIG (Label Information Guided Graph)** — Zhuang et al. (2017): combina aprendizado supervisionado com fatoração matricial, construindo grafos semanticamente coerentes com os rótulos disponíveis (ZHUANG et al., 2017).
- **Matrix Completion Graphs** — Taherkhani et al. (2019): reconstrói conexões ausentes via decomposição de matrizes e aprendizado profundo, capturando relações não observáveis diretamente (TAHERKHANI; KAZEMI; NASRABADI, 2019).
- **IGC (Interactive Graph Construction)** — Chen et al. (2021): incorpora feedback humano no processo de construção, atualizando iterativamente a estrutura da rede (CHEN et al., 2021).
- **Modelos de Partículas (PCC)** — Breve et al. (2011, 2012): utilizam partículas que competem por regiões da rede para propagar rótulos. Versões posteriores tratam *concept drift* e ruído em rótulos (BREVE; ZHAO, 2012; BREVE et al., 2011).
- **JSGFE (Joint Sparse Graph and Flexible Embedding)** — Dornaika & Traboulsi (2019): combina construção de grafos esparsos com aprendizado de embeddings flexíveis (DORNAIKA; TRABOULSI, 2019).

- **Autoencoder-Based Graph Construction** — Kang et al. (2020): utiliza autoencoders com regularização supervisionada para aprender uma matriz de similaridade no espaço latente (KANG et al., 2020).

- **GraphEBM (Energy-Based Graph)** — Chen, Cao & Chang (2020): aprende uma função de energia com uma rede siamesa para definir conectividade entre pares de instâncias (CHEN; CAO; CHANG, 2020).

- **KAOG (K-Associated Optimal Graph)** — Lopes et al. (2009): conecta instâncias da mesma classe com base em pureza estrutural, selecionando o valor ótimo de K de forma adaptativa (LOPES et al., 2009).

- **HL-KAOG (High-Level KAOG)** — Carneiro et al. (2014): complementa o KAOG com métricas de redes complexas (grau médio, agrupamento, assortatividade) para avaliar a compatibilidade de novas instâncias (CARNEIRO et al., 2014).

- **S-kNN (Sequential kNN)** — Vega-Oliveros et al. (2014): evita hubs e distribui conexões uniformemente, construindo grafos mais regulares e estáveis (VEGA-OLIVEROS et al., 2014).

Essas abordagens demonstram que incorporar rótulos desde a construção da rede pode ter impacto mais relevante do que apenas utilizá-los na etapa de propagação. Em particular, métodos supervisionados ou semi-supervisionados tendem a gerar grafos com maior coesão intra-classe e menor ambiguidade, o que contribui diretamente para o sucesso da classificação.

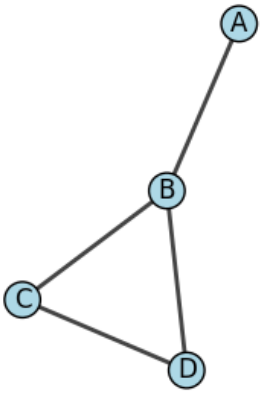
Neste trabalho, serão avaliados os métodos **RGCLI**, **KAOG** e **SNGC**, por representarem diferentes estratégias de construção supervisionada e apresentarem desempenho competitivo na literatura.

2.3 Definições Básicas

A Teoria dos Grafos fornece os fundamentos matemáticos para representar e analisar relações entre entidades por meio de estruturas chamadas grafos. Um grafo G é definido como uma tupla (V, E), onde V representa o conjunto de vértices (ou nós) e E o conjunto de arestas, tal que $E \subseteq V \times V$. Em aplicações de aprendizado de máquina, cada vértice representa uma instância de dado, enquanto uma aresta entre dois vértices indica que existe uma relação (geralmente de similaridade) entre essas instâncias.

Do ponto de vista computacional, grafos podem ser representados por matrizes de adjacência $\mathbf{A}$, nas quais cada elemento $a_{i,j}$ indica a presença (1) ou ausência (0) de uma aresta entre os vértices i e j (GRAFOS. . . , 2012). Para o grafo da Figura 2, temos a seguinte matriz:

Figura 2 – Exemplo de um grafo simples.



Fonte: (Do Autor)

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

2.4 Construção de Grafos

Embora certos sistemas já possuam uma estrutura de rede natural (como redes sociais ou biológicas), na maioria dos cenários de aprendizado de máquina, os dados estão representados como vetores de características. Assim, para aplicar algoritmos baseados em grafos, é necessário construir uma rede a partir desses dados. Essa etapa é chamada de **construção de grafos**.

Vários algoritmos têm sido propostos para essa finalidade, cada um com diferentes critérios de conectividade que resultam em topologias distintas. Essa diversidade influencia diretamente o comportamento dos algoritmos de propagação de rótulos. A seguir, serão descritos os métodos de construção utilizados neste trabalho.

2.4.1 RGCLI: Robust Graph that Considers Labeled Instances

O RGCLI (*Robust Graph that Considers Labeled Instances*) é um método de construção de grafos projetado para cenários de aprendizado semi-supervisionado. Proposto em 2017(BERTON et al., 2017a), o algoritmo visa melhorar a qualidade da propagação de rótulos ao considerar explicitamente a posição dos dados rotulados durante a construção da rede. Em contraste com métodos tradicionais como k-NN, que consideram apenas a distância entre

instâncias, o RGCLI leva em conta a topologia do grafo em relação aos nós com rótulos já conhecidos.

O principal objetivo do RGCLI é gerar uma estrutura de grafo que seja **robusta**, **esparsa** e **informativa**, maximizando a coesão entre dados da mesma classe e minimizando conexões ambíguas entre diferentes classes. Para isso, o método constrói uma rede inicialmente orientada, baseada em um critério que envolve a proximidade com os exemplos rotulados e a distância local, e posteriormente a simetriza, gerando um grafo não orientado(BERTON et al., 2017a).

2.4.2 Descrição do Algoritmo

O processo de construção do grafo pelo RGCLI pode ser resumido em três etapas principais:

1. **Expansão de vizinhança**: para cada instância x_i, é determinada uma vizinhança estendida $N_e(x_i)$ contendo os k_e vizinhos mais próximos com base em uma métrica de distância (geralmente Euclidiana).

2. **Construção das conexões**: a partir de $N_e(x_i)$, são escolhidas k_i conexões com base na soma das distâncias entre x_i e seus vizinhos x_j, e entre cada x_j e o exemplo rotulado mais próximo. Essa heurística prioriza conexões com instâncias próximas a exemplos rotulados, conferindo maior confiança à estrutura da rede.

3. **Formação do grafo**: as conexões formadas para cada instância definem as arestas da rede. Após a construção, a matriz de adjacência é simetrizada, resultando em um grafo não orientado, esparso e robusto, adequado à propagação de rótulos.

2.4.3 Características e Vantagens

O RGCLI apresenta características que o tornam especialmente adequado ao contexto semi-supervisionado:

- **Uso de rótulos na construção**: ao considerar as instâncias rotuladas no momento da conexão, o método fortalece regiões do grafo com maior certeza semântica, evitando conexões entre comunidades de diferentes classes.

- **Preservação da topologia local**: mantém a estrutura local de vizinhança dos dados, mas com foco em conexões que reforçam a coesão intra-classe.

- **Esparsidade**: ao selecionar apenas k_i conexões de uma vizinhança expandida, o grafo gerado é mais leve e eficiente para manipulação computacional.

- **Robustez**: por reduzir as conexões ambíguas (interclasse), o método melhora a confiabilidade da propagação dos rótulos, especialmente quando há poucos exemplos rotulados.

2.4.4 Complexidade Computacional

A complexidade do RGCLI é $O(n \cdot k_e + n \cdot m)$, onde n é o número de instâncias, k_e é o número de vizinhos considerados na vizinhança expandida, e m é o número de instâncias rotuladas. Isso o torna eficiente mesmo em conjuntos de dados com milhares de amostras.

2.4.5 Validação Experimental

O algoritmo RGCLI foi validado empiricamente pelos autores em múltiplos cenários de classificação e segmentação de imagens, utilizando conjuntos de dados como COIL-20, USPS, MNIST e Reuters. Os resultados mostraram desempenho superior em relação a métodos como k-NN, MKNN e B-Matching, principalmente em condições com baixa taxa de rotulagem. Além disso, apresentou menor custo computacional em relação a alternativas baseadas em otimização mais intensiva.

Por essas razões, o RGCLI é utilizado neste trabalho como um dos métodos de construção de grafos a serem avaliados, com foco na sua capacidade de induzir redes que favoreçam a propagação eficiente e precisa dos rótulos(BERTON et al., 2017b).

2.4.6 K-Associated Optimal Graph (KAOG)

O *K-Associated Optimal Graph* (KAOG) é um método de construção de grafos originalmente proposto para tarefas de classificação supervisionada, mas que pode ser adaptado ao cenário semi-supervisionado por sua habilidade de gerar estruturas topológicas altamente representativas da distribuição das classes nos dados.

O KAOG é composto por duas etapas principais: (i) a construção de múltiplos grafos K-associados e (ii) a composição de um grafo ótimo a partir dos melhores componentes (subgrafos) extraídos desses diferentes grafos (CARNEIRO et al., 2014; LOPES et al., 2009).

2.4.7 Construção de Grafos K-associados

Dado um conjunto de dados rotulados, o grafo K-associado é construído conectando cada instância apenas a seus K vizinhos mais próximos **da mesma classe**. Ou seja, para cada instância v_i, são criadas arestas para os vizinhos v_j tais que $v_j \in \text{knn}(v_i)$ e $\text{classe}(v_j) = \text{classe}(v_i)$. O grafo gerado possui componentes (subgrafos conectados) cujos vértices compartilham o mesmo rótulo.

2.4.8 Medida de Pureza dos Componentes

A pureza de um componente representa o grau de conectividade entre vértices de mesma classe, e serve como critério para selecionar as melhores componentes ao longo das iterações. Formalmente, define-se a pureza ϕ como:

$$\phi = \frac{\langle g \rangle}{2K} \tag{1}$$

onde $\langle g \rangle$ é o grau médio dos vértices do componente e K é o número de vizinhos usados na construção do grafo. Essa métrica varia entre 0 e 1, sendo que valores próximos a 1 indicam alta densidade de conexões internas entre vértices da mesma classe.

2.4.9 Grafo Ótimo

A construção do *grafo ótimo* consiste em iterar sobre diferentes valores de K e, a cada passo, reter os componentes que maximizam uma medida de qualidade estrutural. Essa medida pode ser a própria pureza ϕ, ou uma métrica ponderada W que considera também o tamanho do componente:

$$W = \frac{\sum_{i=1}^{N} g_i}{2K} \quad \text{com } \langle g \rangle > K \tag{2}$$

A ideia é construir um grafo final contendo as melhores componentes encontradas em diferentes valores de K, permitindo que cada região do espaço de atributos seja representada por um valor de K que maximize sua conectividade interna. Isso resulta em uma representação topológica robusta e adaptada à geometria das classes nos dados.

2.4.10 Classificação com KAOG

No cenário semi-supervisionado, as componentes extraídas do grafo KAOG podem servir como núcleos de confiança para propagação de rótulos, por exemplo por meio de algoritmos como o *Label Propagation*. Componentes altamente puras atuam como regiões confiáveis para inferência de rótulos em dados não rotulados. Alternativamente, abordagens baseadas em classificação bayesiana ou em classificação de alto nível (*high-level classification*) podem ser aplicadas, explorando tanto atributos físicos quanto padrões topológicos (CARNEIRO et al., 2014).

2.4.11 Vantagens

- **Adaptabilidade topológica**: cada componente pode ser construída com um valor diferente de K, otimizando sua estrutura local.
- **Preservação estrutural**: os subgrafos mantêm padrões topológicos das classes, mesmo em contextos com ruído ou formas complexas.
- **Integração com outras técnicas**: o KAOG pode ser utilizado como base para algoritmos de propagação, classificadores bayesianos ou métodos híbridos.

2.4.12 Limitações

Apesar de suas vantagens, o KAOG pode apresentar custo computacional elevado, especialmente pela necessidade de calcular múltiplos grafos e componentes. Além disso, a qualidade das componentes pode ser impactada em cenários com escassez extrema de dados rotulados.

2.4.13 Aplicação neste trabalho

Neste trabalho, o KAOG será utilizado como um dos métodos de construção de grafos no cenário de aprendizado semi-supervisionado. Serão analisadas as propriedades topológicas dos grafos gerados, bem como o desempenho de algoritmos de propagação de rótulos sobre eles.

2.4.14 SNGC: Supervised Neighborhood Graph Construction

O *Supervised Neighborhood Graph Construction* (SNGC) é um método de construção de grafos proposto por Rohban e Rabiee em 2012(ROHBAN; RABIEE, 2012), especificamente voltado para cenários de aprendizado semi-supervisionado. Seu objetivo é melhorar a qualidade dos grafos utilizados para propagação de rótulos, incorporando diretamente informações supervisionadas (rótulos parciais) no processo de construção da vizinhança. Diferentemente de abordagens puramente geométricas como o k-NN, o SNGC busca criar conexões que respeitem tanto a estrutura local dos dados quanto a semântica induzida pelos exemplos rotulados.

2.4.15 Intuição do Método

O SNGC parte da premissa de que a conectividade em um grafo para aprendizado semi-supervisionado deve refletir a probabilidade de que duas instâncias pertençam à mesma classe. Assim, o método visa penalizar conexões entre instâncias de classes diferentes e favorecer conexões entre instâncias da mesma classe, mesmo que estas não estejam necessariamente próximas no espaço de atributos. Para isso, propõe-se o aprendizado de uma função de vizinhança supervisionada baseada na probabilidade de erro de ligação entre pares de dados.

2.4.16 Etapas do Algoritmo

O processo de construção do grafo supervisionado no SNGC pode ser descrito em três etapas principais:

1. **Geração do grafo base (k_0-NN)**: inicia-se construindo um grafo inicial G_0 a partir do k_0-NN simétrico tradicional, onde cada ponto é conectado aos seus k_0 vizinhos mais próximos com base em distância (usualmente Euclidiana). Esse grafo define os pares candidatos a serem avaliados.

2. **Treinamento de um classificador de links**: a partir dos exemplos rotulados disponíveis, são formados pares de pontos (x_i, x_j) com seus respectivos vetores de diferença $(x_j - x_i)$ e um rótulo binário indicando se pertencem à mesma classe (0) ou não (1). Um classificador supervisionado (geralmente um SVM com kernel RBF) é treinado para estimar a probabilidade de erro de ligação $\Pr(y_i \neq y_j)$.

3. **Construção do grafo supervisionado**: para cada ponto x_i, os vizinhos candidatos x_j (vindos do grafo G_0) são avaliados com o classificador treinado. São mantidos os k vizinhos com maior confiança de serem da mesma classe, definida como $1 - \Pr(y_i \neq y_j)$, desde que essa confiança ultrapasse um limiar mínimo θ (geralmente 0.75). O resultado é um grafo esparso, supervisionado, e adaptado à estrutura das classes.

2.4.17 Características e Vantagens

O SNGC apresenta várias vantagens no contexto semi-supervisionado:

- **Incorporação explícita de supervisão**: usa rótulos parciais para guiar a construção do grafo, evitando conexões entre instâncias de classes distintas.
- **Aprendizado de função de vizinhança**: estima diretamente a probabilidade de ligação correta entre pares, ao invés de depender unicamente de distância.
- **Esparsidade e seletividade**: mantém apenas os vizinhos mais confiáveis, reduzindo ruídos e conexões ambíguas.
- **Flexibilidade**: pode ser combinado com diferentes classificadores de links e ajustado com limiares de confiança para regular a densidade do grafo.

2.4.18 Aplicação neste trabalho

Neste trabalho, o SNGC é utilizado como um dos métodos de construção de grafos no contexto de propagação de rótulos. Sua característica de selecionar conexões com alta confiança supervisionada o torna especialmente interessante para cenários com poucos rótulos disponíveis, permitindo avaliar a eficácia de redes supervisionadas frente a alternativas não supervisionadas como o k-NN.

2.5 Algoritmos Clássicos de Classificação com Grafos

No contexto do aprendizado semi-supervisionado baseado em grafos, a etapa de propagação de rótulos é responsável por estender a informação dos exemplos rotulados para os demais nós da rede. Após a construção do grafo de similaridade entre instâncias, os algoritmos de propagação assumem que vértices conectados — ou seja, amostras semelhantes — devem pertencer

à mesma classe. Essa suposição é conhecida como **hipótese da suavidade** (*smoothness assumption*), segundo a qual os rótulos devem variar de forma suave ao longo da estrutura do grafo.

Dentre os algoritmos mais clássicos e amplamente utilizados na literatura estão:

- **Label Propagation**: este algoritmo inicializa os rótulos apenas nos nós rotulados e propaga essa informação para os não rotulados com base na conectividade do grafo. A cada iteração, cada nó não rotulado adota o rótulo mais frequente entre seus vizinhos. Os rótulos dos nós inicialmente rotulados permanecem fixos durante toda a execução, funcionando como fontes de informação. O processo itera até que os rótulos se estabilizem, ou seja, não haja mais mudanças entre as iterações.(RAMOS et al., 2020)(SANTOS, 2023) (DATACAMP, 2023)

- **Label Spreading**: considerado uma generalização do Label Propagation, o Label Spreading introduz uma matriz de transição suavizada, baseada na similaridade entre instâncias. Diferente do anterior, esse algoritmo permite a atualização dos rótulos dos nós rotulados, embora com menor influência que os não rotulados, controlando esse efeito por meio de um parâmetro de regularização. O uso de uma matriz simétrica e suavizada tende a tornar a propagação mais estável, sendo especialmente útil em bases com ruído ou sobreposição entre classes.(RAMOS, 2020)(scikit-learn developers, 2024)

Ambos os algoritmos compartilham o princípio de que a topologia do grafo exerce forte influência sobre os resultados da classificação. Grafos mal construídos — com ligações indevidas entre diferentes classes — podem comprometer a qualidade da propagação, resultando em erros acumulativos. Por isso, é fundamental garantir que o grafo represente corretamente a estrutura semântica dos dados, refletindo relações reais de similaridade entre instâncias.

Além desses dois, existem também outras variantes na literatura, como métodos baseados em funções harmônicas e técnicas que combinam propagação com regularização global. No entanto, Label Propagation e Label Spreading permanecem como os mais consolidados e utilizados, sendo ambos implementados nativamente em bibliotecas como a Scikit-Learn.

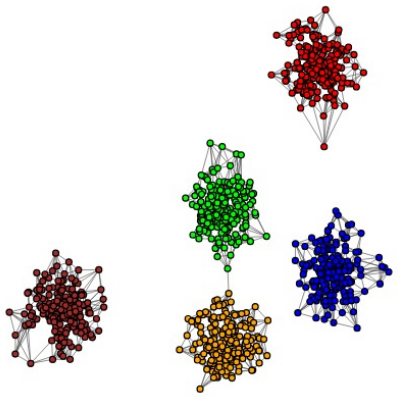
2.6 Estrutura Comunitária e Classificação em Grafos

Embora a detecção de comunidades seja um tema clássico do aprendizado não supervisionado, o conceito de **estrutura comunitária** também é relevante no contexto do aprendizado semi-supervisionado com grafos. Em grafos, comunidades correspondem a subconjuntos de vértices que são densamente conectados entre si e relativamente pouco conectados com o restante do grafo. Em termos práticos, tais comunidades geralmente refletem grupos de

instâncias que compartilham características similares — o que, em muitos casos, corresponde às classes verdadeiras dos dados.(SANTOS, 2016)(GUERREIRO, 2019)

Nos algoritmos de propagação de rótulos, a presença de uma estrutura comunitária bem definida favorece a correta difusão da informação. Vértices localizados em uma mesma comunidade tendem a receber o mesmo rótulo de forma mais consistente e confiável, devido à maior coesão local. Isso contribui diretamente para a precisão do processo classificatório.

Figura 3 – Grafo com estrutura comunitária favorecendo a propagação de rótulos.



Fonte: (Do Autor)

A Figura 3 ilustra um exemplo de grafo dividido em comunidades, com vértices conectados fortemente dentro dos grupos. A propagação de rótulos, quando realizada sobre esse tipo de estrutura, tende a preservar a coerência da informação inicial, minimizando o risco de que instâncias de diferentes classes compartilhem rótulos incorretos.

Assim, ao avaliar algoritmos de construção de grafos neste trabalho, também será considerado como cada método influencia na formação (ou deterioração) dessa estrutura comunitária, impactando diretamente no sucesso dos algoritmos de classificação semi-supervisionada.

3 METODOLOGIA

Esta seção descreve os métodos e técnicas, assim como os dados e ferramentas que foram utilizados para conduzir os experimentos. Todos os algoritmos utilizados e implementados utilizaram a linguagem de programação Python.

Resumidamente, os experimentos foram desenvolvidos por meio das seguintes etapas de elaboração do estudo:

- Escolha dos algoritmos
- Geração do conjunto de dados

- Construção de redes
- Realização dos experimentos
- Avaliação dos resultados

3.1 Escolha dos Algoritmos

Os algoritmos selecionados para a realização dos experimentos foram detalhados nas seções anteriores, como parte da Revisão Bibliográfica. A seguir estão descritos os principais parâmetros para implementação de cada um dos algoritmos. Esses parâmetros serão sistematicamente variados ao longo dos experimentos para avaliar sua influência no desempenho da classificação semi-supervisionada:

RGCLI (Robust Graph that Considers Labeled Instances)

- k_e — número de vizinhos considerados na construção do grafo kNN simétrico inicial.
- k_i — número de vizinhos confiáveis selecionados com base em similaridade e proximidade a instâncias rotuladas.

KAOG (K-Associated Optimal Graph)

- $K_{\max}$ — valor máximo de vizinhança avaliado durante a busca do valor ótimo de K.
- Métrica de pureza (ϕ) — utilizada para selecionar a configuração ótima de K para cada subgrafo rotulado.

SNGC (Supervised Neighborhood Graph Construction)

- k — número final de vizinhos confiáveis conectados a cada ponto no grafo supervisionado.
- k_0 — número de vizinhos no grafo inicial simétrico usado como hipótese.
- γ — parâmetro do kernel RBF do classificador SVM utilizado para avaliar a confiabilidade das conexões.
- $\text{conf}_{\min}$ — limiar mínimo de confiança (probabilidade de duas instâncias pertencerem à mesma classe) para aceitação de uma aresta.

Os valores desses parâmetros serão explorados em conjunto com diferentes porcentagens de instâncias rotuladas para cada base de dados, com o objetivo de compreender como cada método responde à variação de supervisão e sensibilidade a ajustes na construção do grafo.

3.2 Conjunto de Dados

Neste trabalho, serão utilizadas quatro bases de dados sintéticas, geradas com o auxílio da biblioteca Scikit-Learn (Iris, Wine, Breast Cancer e Diabetes), além de um conjunto de dados textuais (20 Newsgroups), amplamente empregado em tarefas de classificação de textos e processamento de linguagem natural.

3.2.1 Iris

O dataset Iris é um clássico conjunto de dados multivariados introduzido por Ronald Fisher em 1936. Contém 150 instâncias distribuídas em três classes de flores Iris (Setosa, Versicolour e Virginica), com 50 amostras por classe. Cada amostra é descrita por quatro características: comprimento e largura da sépala, e comprimento e largura da pétala (em cm). Por sua simplicidade e clara separabilidade entre classes, é frequentemente utilizado para tarefas de classificação.

3.2.2 Wine

O dataset Wine consiste em 178 amostras de vinhos provenientes de três diferentes cultivares na região da Itália. Cada amostra é caracterizada por 13 atributos químicos (como teor de álcool, ácido málico, concentração de flavonoides, etc.). Este conjunto é frequentemente utilizado para problemas de classificação multiclasse, apresentando um desafio moderado devido à sobreposição entre algumas classes.

3.2.3 Breast Cancer

O dataset Breast Cancer (Wisconsin Diagnostic Breast Cancer - WDBC) contém 569 amostras de características de núcleos celulares presentes em imagens de massa mamária, classificadas como malignas ou benignas. Cada instância possui 30 atributos numéricos derivados de características como raio, textura, perímetro e área dos núcleos celulares. Este conjunto é amplamente utilizado para avaliação de classificadores binários na área médica.

3.2.4 Diabetes

O dataset Diabetes (também conhecido como "diabetes dataset"do Scikit-Learn) é um conjunto de dados de regressão que contém 442 amostras de pacientes diabéticos. Cada instância possui 10 atributos fisiológicos (idade, sexo, índice de massa corporal, pressão arterial, etc.) e uma medida quantitativa da progressão da doença um ano após a linha de base.

3.2.5 20 Newsgroups

O dataset *20 Newsgroups* é uma coleção de aproximadamente **20.000 documentos textais**, divididos em **20 categorias temáticas**, como `sci.space`, `rec.sports.baseball`, `comp.graphics` e `talk.politics.mideast`. Originalmente coletado por Ken Lang em 1995, esse conjunto é frequentemente utilizado para avaliar algoritmos de aprendizado de máquina em tarefas de classificação de textos, extração de tópicos e modelagem de linguagem.

Cada documento consiste em um post de um grupo de discussão (*newsgroup*), contendo texto puro, cabeçalhos e metadados (como autor e assunto). O dataset pode ser carregado em diferentes configurações, incluindo versões balanceadas (com número igual de documentos por classe) ou subconjuntos específicos para tarefas binárias ou multiclasse.

Devido à sua natureza realista e diversidade de tópicos, o *20 Newsgroups* é um *benchmark* relevante para avaliar a capacidade de modelos em lidar com dados textuais estruturados e não estruturados, sendo particularmente útil para testar técnicas de pré-processamento, vetorização (como *TF-IDF* ou *word embeddings*) e classificação baseada em conteúdo.

3.3 Ferramentas Utilizadas

3.3.1 Linguagem de Programação

Todo o desenvolvimento será realizado em **Python**, linguagem de programação de alto nível, interpretada, de código aberto e com forte presença nas áreas de ciência de dados, inteligência artificial e aprendizado de máquina. Sua sintaxe simples e legibilidade favorecem o desenvolvimento rápido de protótipos e a análise exploratória de dados, além de contar com uma vasta comunidade ativa e suporte robusto para bibliotecas científicas.

3.3.2 Bibliotecas

Diversas bibliotecas do ecossistema Python serão utilizadas neste trabalho:

- **Scikit-Learn**: biblioteca de código aberto voltada ao aprendizado de máquina, com suporte a algoritmos supervisionados, não supervisionados e semi-supervisionados. Também oferece ferramentas para pré-processamento, geração de dados sintéticos, validação cruzada e métricas de avaliação (PEDREGOSA et al., 2011). Depende de bibliotecas como NumPy[1], SciPy[2] e Matplotlib[3] para operações numéricas e visualização.
- **Pandas**: biblioteca para manipulação e análise de dados estruturados, com foco em tabelas (DataFrames). Permite importar dados de múltiplas fontes (CSV, Excel, SQL), realizar operações vetoriais, agregações, filtros e transformações com alta performance.

[1] <https://numpy.org/>
[2] <https://scipy.org/>
[3] <https://matplotlib.org/>

- **NumPy**: oferece suporte a arrays e operações matemáticas vetorizadas de alto desempenho. É a base para diversas outras bibliotecas científicas do Python.
- **SciPy**: biblioteca complementar ao NumPy que fornece funcionalidades matemáticas avançadas, como álgebra linear, estatística, interpolação e integração numérica.
- **Matplotlib**: utilizada para visualização gráfica de dados. Permite criar desde gráficos simples até visualizações mais elaboradas, como mapas de calor, matrizes de confusão e visualizações de redes.
- **Igraph e NetworkX**: bibliotecas especializadas em análise de grafos. Permitem a criação, manipulação e visualização de redes, além da execução de algoritmos clássicos como extração de componentes conexos, medidas de centralidade, distâncias e caminhamentos.

3.4 Metodologia Experimental

O trabalho tem natureza experimental e quantitativa, pois visa comparar métodos de construção de grafos no contexto do aprendizado semi-supervisionado. O fluxo geral da metodologia é descrito abaixo:

1. Os dados são carregados e pré-processados.
2. Uma fração dos dados é rotulada (as porcentagens 80% 70%, 60% e 50% serão testadas).
3. Um grafo é construído a partir dos dados com base em um dos métodos propostos: RGCLI, KAOG ou SNGC.
4. O algoritmo de propagação de rótulos *Label Propagation* é aplicado sobre o grafo construído.
5. O desempenho é avaliado por meio da comparação entre os rótulos preditos e os rótulos reais dos dados não rotulados.

Cada combinação (grafo + algoritmo de propagação + proporção de dados rotulados) será testada múltiplas vezes para garantir estabilidade estatística e média dos resultados.

3.5 Experimentos

Os experimentos realizados têm como objetivo avaliar o impacto dos diferentes métodos de construção de grafos (RGCLI, KAOG e SNGC) sobre o desempenho de algoritmos de aprendizado semi-supervisionado, em particular o *Label Propagation*.

O protocolo experimental adotado consiste nas seguintes etapas:

- Os dados são carregados e pré-processados. Os conjuntos utilizados são: Iris, Wine, Breast Cancer, Diabetes e 20 Newsgroups.

- Uma fração dos dados é rotulada de forma estratificada, sendo testadas as porcentagens de 80%, 70%, 60% e 50% de amostras rotuladas (isto é, 20%, 30%, 40% e 50% de amostras não rotuladas). Os índices rotulados são mantidos fixos por semente aleatória para permitir reprodutibilidade.

- Um grafo é construído a partir dos dados utilizando um dos três métodos selecionados:

  - **RGCLI:** para cada porcentagem de dados rotulados, fixou-se o valor de $k_e = 10$ e variou-se o parâmetro k_i no intervalo de 2 a 8. O objetivo foi avaliar o impacto da densidade das conexões confiáveis no desempenho da propagação de rótulos. Em experimentos complementares, também foi avaliada a sensibilidade ao parâmetro k_e, mantendo k_i fixo.
  - **KAOG:** para cada porcentagem de dados rotulados, o valor de $K_{\max}$ foi variado de 3 a 15. Para cada valor de K, o algoritmo constrói subgrafos locais e avalia a métrica de pureza ϕ, selecionando automaticamente a vizinhança ótima para cada componente rotulada. O foco do experimento está em observar como a diversidade de topologias geradas por diferentes K influencia a classificação.
  - **SNGC:** para cada porcentagem de dados rotulados, foi adotado um valor fixo para $k_0 = 10$ (grafo inicial simétrico), enquanto o número final de conexões confiáveis k foi variado de 2 a 8. Além disso, o parâmetro de confiança mínima $\text{conf}_{\min}$ foi testado com os valores 0.6, 0.7 e 0.8.
  - **kNN (baseline tradicional)**: como forma de comparação, também foi utilizado o método tradicional de construção de grafos baseado em k-Nearest Neighbors, sem uso de rótulos. Para cada ponto, são conectados os k vizinhos mais próximos com base na distância euclidiana. O valor de k foi variado de 5 a 15. Este método serve como referência não supervisionada para avaliar o ganho obtido ao incorporar informações de rótulos na construção do grafo (MAIER; HEIN; LUXBURG, 2007).

- O algoritmo *Label Propagation* é aplicado sobre o grafo resultante, propagando os rótulos conhecidos até os nós não rotulados.

- O desempenho da classificação é avaliado comparando os rótulos preditos com os reais, utilizando métricas como acurácia, para os dados originalmente não rotulados. Os experimentos são repetidos múltiplas vezes para obter médias e desvios padrão.

Considerações sobre o Algoritmo KAOG

O algoritmo original do *K-Associated Optimal Graph* (KAOG), como proposto na literatura, foi concebido para cenários supervisionados, conectando apenas amostras com rótulos e restringindo as conexões a pontos da mesma classe. Tal abordagem tende a gerar componentes desconectadas quando aplicado ao cenário semi-supervisionado, sobretudo em bases com proporção significativa de dados não rotulados. Essa desconexão compromete a propagação de rótulos, impedindo que muitos nós recebam influências das instâncias rotuladas.

Diante disso, optou-se por uma versão adaptada do algoritmo, descrita a seguir:

- **Critério de remoção de arestas suavizado**: enquanto o KAOG original remove todas as conexões entre classes distintas, a versão adotada remove apenas arestas entre **duas amostras rotuladas de classes diferentes**. Conexões entre amostras não rotuladas (ou entre rotulada e não rotulada) são preservadas, o que permite maior conectividade no grafo.

- **Componentes analisadas**: a versão original considera apenas componentes conexas compostas inteiramente por amostras rotuladas. Nesta adaptação, todos os componentes do grafo são considerados, mas apenas aqueles que contêm uma proporção dominante de um mesmo rótulo (isto é, cujas instâncias rotuladas pertencem à mesma classe) são retidos como subgrafos válidos.

- **Inclusão de instâncias não rotuladas**: como consequência da mudança anterior, o grafo final pode incluir instâncias não rotuladas desde que estas estejam conectadas a subgrafos majoritariamente rotulados. Isso melhora a cobertura da propagação de rótulos, mantendo a ideia de pureza estrutural baseada na métrica ϕ.

- **Condições de substituição de subgrafos**: foi mantido o critério original em que, para cada classe, componentes maiores (ou com maior pureza ϕ) substituem subgrafos anteriores.

Essas modificações permitiram que o KAOG operasse de forma mais eficaz em contextos semi-supervisionados, garantindo conectividade adequada para a propagação de rótulos sem violar os princípios de separação por classe que fundamentam o método.

Considerações sobre o Algoritmo SNGC

O algoritmo *Robust Graph Construction that Considers Labeled Instances* (SNGC) foi originalmente proposto para construir grafos confiáveis com base em classificadores treinados sobre pares rotulados. No entanto, sua implementação original apresenta alto custo computacional, especialmente em conjuntos de dados com grande número de amostras rotuladas, dado que considera todos os pares possíveis para treinamento do classificador.

Para viabilizar a execução do SNGC em todos os experimentos do presente trabalho, foi utilizada uma versão otimizada da implementação, com foco em desempenho computacional sem comprometer os princípios fundamentais do método.

As principais adaptações foram:

- **Limite de amostras de pares rotulados**: ao invés de utilizar todos os pares possíveis entre instâncias rotuladas (custo $O(n^2)$), foi adotado um limite máximo de 3000 pares. Caso o número total de pares ultrapasse esse limite, uma amostragem aleatória é realizada. Isso reduz drasticamente o tempo de treinamento do classificador SVM.

- **Uso de calibrador de probabilidades**: o classificador SVM original, baseado em kernel RBF, foi encapsulado com um `CalibratedClassifierCV`, que aplica calibração por *sigmoid* (Platt scaling). Essa abordagem estabiliza a estimativa das probabilidades associadas às conexões confiáveis, mantendo o controle de confiança do método original.

- **Construção vetorial do grafo inicial k_0-NN**: ao invés de calcular distâncias ponto a ponto em laços separados, a matriz de distâncias é computada de forma vetorizada com a função `pairwise_distances`, seguida da construção direta da matriz de adjacência simétrica.

- **Seleção eficiente das conexões finais**: para cada instância, apenas os k vizinhos com maior confiança (entre os que superam o limiar conf_{min}) são mantidos, conforme o critério original do SNGC. A ordenação e filtragem são feitas de forma vetorizada.

Tais modificações mantêm os princípios centrais do SNGC — ou seja, a utilização de pares rotulados para aprendizado de confiabilidade entre conexões —, mas tornam sua execução prática e escalável para múltiplas repetições experimentais. Essa adaptação se mostrou essencial para garantir a viabilidade da comparação entre métodos no escopo deste trabalho.

Considerações sobre o conjunto `Diabetes`:

O conjunto de dados `Diabetes`, fornecido pelo *scikit-learn*, é tradicionalmente utilizado para tarefas de regressão, onde a variável alvo representa uma medida contínua associada à progressão da doença em pacientes(PEDREGOSA et al., 2011).

Como o presente trabalho tem como foco algoritmos de aprendizado semi-supervisionado voltados à **classificação**, foi necessário transformar o problema original em uma tarefa de classificação binária. Para isso, aplicou-se uma discretização simples: as amostras foram rotuladas com 1 se o valor do alvo original estivesse acima da mediana da distribuição, e com 0 caso contrário.

Essa transformação garante que o conjunto resultante seja compatível com os métodos de propagação de rótulos baseados em grafos, além de preservar a natureza relativa dos dados

originais. A nova variável binária representa, de forma simplificada, pacientes com maior ou menor risco relativo de progressão da doença.

Considerações especiais sobre o conjunto 20 Newsgroups:
Para este conjunto textual, foi utilizado um subconjunto contendo 5 categorias específicas:

- `alt.atheism`
- `talk.religion.misc`
- `comp.graphics`
- `soc.religion.christian`
- `sci.space`

Essas categorias foram escolhidas por apresentarem diversidade temática, o que permite avaliar a capacidade dos métodos de construção de grafos em contextos com fronteiras interclasse mais complexas.

A representação textual foi feita por meio de vetores *TF-IDF* (Term Frequency–Inverse Document Frequency), extraídos com a limitação de até 20 termos mais frequentes. Técnicas adicionais de pré-processamento, lematização ou stemming não foram aplicadas neste trabalho, de forma a preservar a estrutura original do vocabulário para análise da conectividade semântica entre documentos.

3.6 Avaliação dos Resultados

Para avaliar a qualidade das classificações obtidas, foram utilizadas métricas padrão na literatura de Classificação(FERRI; HERNáNDEZ-ORTEGA; HERNáNDEZ-ORALLO, 2001; TAN; STEINBACH; KUMAR, 2009), juntamente com um teste estatístico:

- **Acurácia**: mede a proporção de rótulos corretamente classificados.
- **F1-score**: métrica harmônica entre precisão e revocação, útil especialmente em bases com classes desbalanceadas.
- **AUC (Area Under the Curve)**: representa a área sob a curva ROC (Receiver Operating Characteristic), indicando a capacidade do modelo de distinguir entre as classes. É particularmente útil em bases com classes desbalanceadas. Para problemas com mais de duas classes, foi utilizada a estratégia *one-vs-rest* com a função `roc_auc_score` da biblioteca Scikit-Learn.

3.6.1 Análise Estatística dos Resultados

Para além da análise descritiva das métricas de desempenho (acurácia, F1-score e AUC), é fundamental avaliar se as diferenças observadas entre os métodos de construção de grafos são estatisticamente significativas. Para isso, recorreu-se a testes estatísticos formais, capazes de verificar se os resultados obtidos se devem ao acaso ou refletem vantagens reais dos algoritmos.

O processo de análise estatística seguiu diretrizes clássicas da literatura de comparação entre algoritmos de aprendizado de máquina, conforme sugerido por Demšar (DEMšAR, 2006). Inicialmente, são aplicados testes de **normalidade** (*Shapiro-Wilk*) para verificar se os resultados seguem uma distribuição normal, e testes de **homocedasticidade** (isto é, igualdade de variâncias) — como o *teste de Bartlett* — para assegurar que os métodos têm variâncias comparáveis(SHAPIRO; WILK, 1965; BARTLETT, 1937).

Quando os dados são considerados normais e homocedásticos, aplica-se a **ANOVA de medidas repetidas** (*Repeated Measures ANOVA*), um teste paramétrico apropriado para comparar o desempenho de múltiplos métodos sobre os mesmos conjuntos de dados. Se a ANOVA indicar diferenças significativas entre os métodos, é realizado um **teste post hoc de Tukey HSD** (*Honestly Significant Difference*) para identificar quais pares de algoritmos apresentam diferenças estatisticamente relevantes(FIELD, 2018; TUKEY, 1953).

Para conduzir essa análise de forma sistemática e reprodutível, utilizou-se a biblioteca `Autorank` (HERBOLD, 2020), que automatiza a seleção e aplicação dos testes adequados com base nas propriedades dos dados. A ferramenta realiza também correções para comparações múltiplas (como a correção de Bonferroni) e gera visualizações e relatórios prontos para inclusão em documentos LaTeX.

Esta abordagem foi aplicada sobre os resultados consolidados dos experimentos, considerando os algoritmos **RGCLI**, **KAOG**, **SNGC** e **kNN**, avaliados em cinco conjuntos de dados distintos. As métricas analisadas foram **acurácia**, **F1-score** e **AUC**. Os resultados da análise estatística são discutidos na próxima subseção.

Tabela 1 – Resumo do protocolo experimental adotado

Etapa	**Descrição**
Conjuntos de dados	Iris, Wine, Breast Cancer, Diabetes, 20 Newsgroups (5 categorias específicas)
Porcentagens de rotulagem	80%, 70%, 60% e 50% de amostras rotuladas (estratificadas)
Semente aleatória	Usada para fixar os índices rotulados e garantir reprodutibilidade
Métodos de construção de grafos	• **RGCLI:** $k_e = 10$ fixo; $k_i \in \{2, 3, \ldots, 8\}$ • **KAOG:** $K_{\max} \in \{3, 4, \ldots, 15\}$ • **SNGC:** $k_0 = 10$ fixo; $k \in \{2, \ldots, 8\}$; $\text{conf}_{\min} \in \{0.6, 0.7, 0.8\}$; $\gamma =$ `'scale'` • **kNN (baseline):** $k \in \{5, \ldots, 15\}$
Classificador	*Label Propagation*, aplicado sobre o grafo construído
Avaliação	Acurácia, F1-score e AUC-ROC (estratégia *one-vs-rest*)

4 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Nesta seção, são analisados os resultados obtidos nos experimentos descritos anteriormente, com base nas métricas de *accuracy*, *f1-score* e *AUC*. A discussão está organizada de forma a destacar o comportamento dos métodos frente à variação de seus parâmetros, a influência da proporção de dados rotulados e a comparação estatística entre os algoritmos.

4.1 Desempenho Geral por Método

Considerando todos os datasets e diferentes proporções de rótulos, observa-se que o método **kNN (baseline)** apresentou desempenho competitivo e consistente em todas as métricas, refletindo a robustez do modelo mesmo sem utilizar informações supervisionadas na construção do grafo. O método **RGCLI** demonstrou desempenho semelhante ao kNN, com destaque para a métrica de *AUC*, especialmente em bases com menos sobreposição entre classes.

O **SNGC**, embora mais custoso computacionalmente, apresentou boa capacidade discriminativa, com destaque em bases de maior complexidade, como **20ng**. O **KAOG**, por sua vez, obteve desempenho inferior, com acurácia e f1-score consistentemente menores, fato possivelmente relacionado à sua dependência exclusiva de informações rotuladas para conexões.

Além disso, ao comparar as médias gerais entre os métodos, o kNN obteve as maiores médias de acurácia (0,742), f1-score (0,727) e AUC (0,866), seguido de perto pelo RGCLI, que apresentou desempenho competitivo e estável. Já o SNGC demonstrou boa performance em bases mais complexas, mas com maior variabilidade. O KAOG, por outro lado, foi consistentemente o método de menor desempenho, refletindo limitações estruturais em sua construção de grafo.

Base Iris

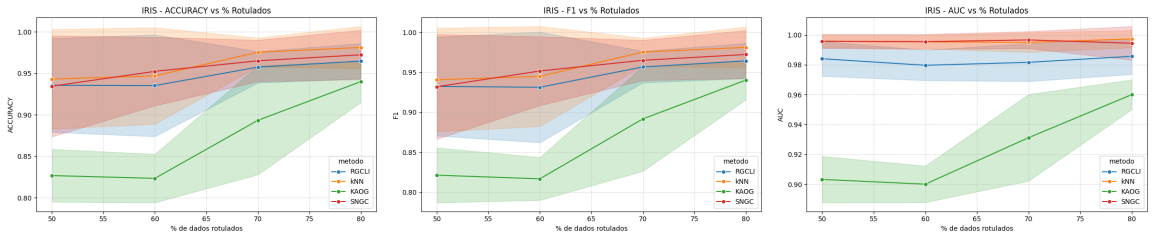


Figura 4 – Curvas de acurácia, F1-score e AUC para os métodos na base **Iris**, em função da proporção de rótulos disponíveis.

Base Wine

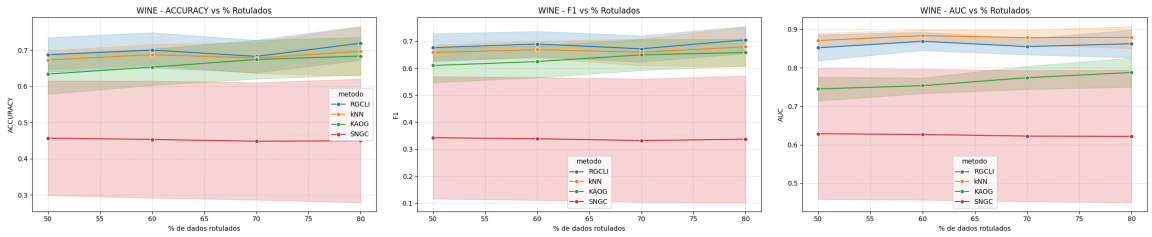


Figura 5 – Curvas de acurácia, F1-score e AUC para os métodos na base **Wine**, em função da proporção de rótulos disponíveis.

Base Breast Cancer

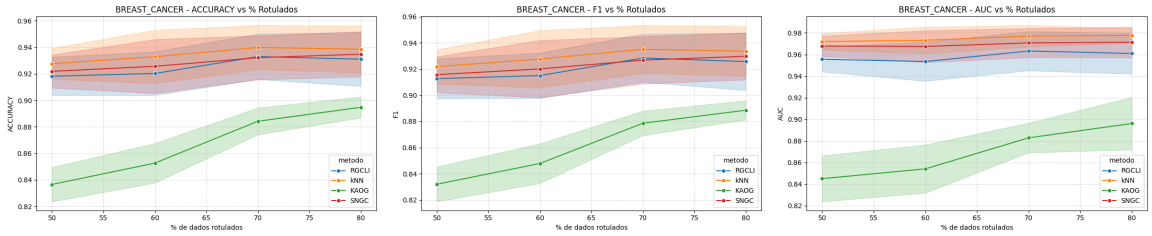


Figura 6 – Curvas de acurácia, F1-score e AUC para os métodos na base **Breast Cancer**, em função da proporção de rótulos disponíveis.

Base Diabetes

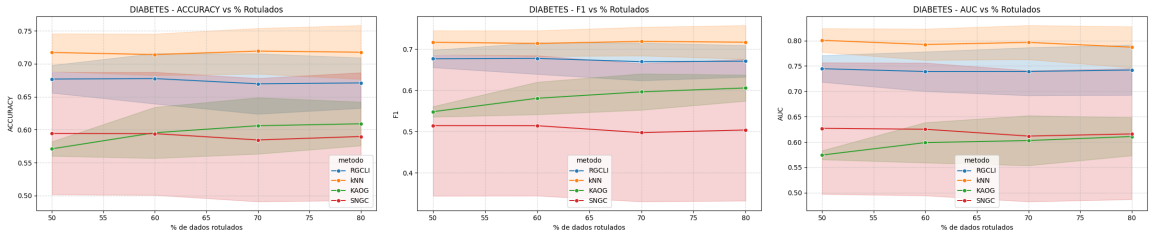


Figura 7 – Curvas de acurácia, F1-score e AUC para os métodos na base **Diabetes**, em função da proporção de rótulos disponíveis.

Base 20 Newsgroups

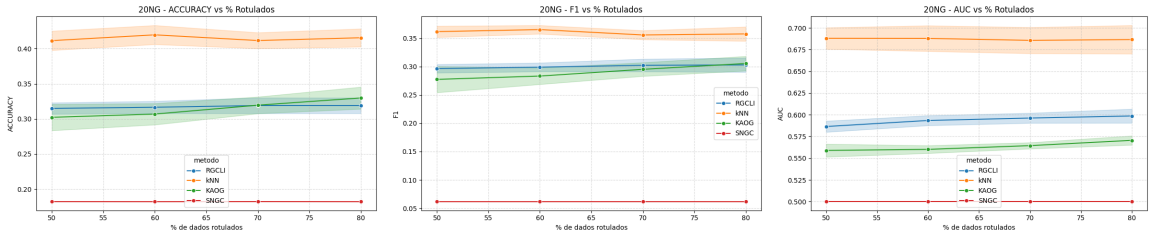


Figura 8 – Curvas de acurácia, F1-score e AUC para os métodos na base **20 Newsgroups**, em função da proporção de rótulos disponíveis.

4.2 Impacto da Porcentagem de Rótulos

Como esperado, o desempenho dos métodos tende a crescer com o aumento da porcentagem de amostras rotuladas. Esse comportamento é mais evidente nos métodos RGCLI e SNGC, que se beneficiam diretamente das informações supervisionadas para a construção do grafo. No KAOG, o impacto é menos pronunciado, devido à exclusão de vários nós não rotulados das conexões finais.

Em particular, observou-se que o kNN apresenta maior estabilidade com variações de rotulagem, enquanto métodos como o SNGC sofrem mais com a escassez de dados rotulados, o que reduz a eficácia da etapa supervisionada.

Base Iris

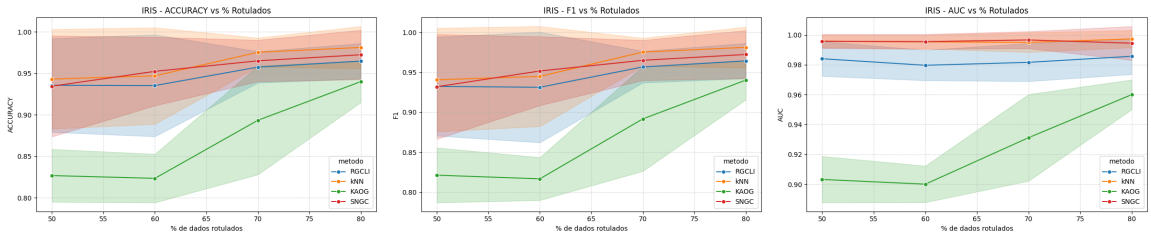


Figura 9 – Desempenho dos métodos na base **Iris** em função da porcentagem de dados rotulados. As curvas mostram média e desvio padrão ao longo das repetições.

Base Wine

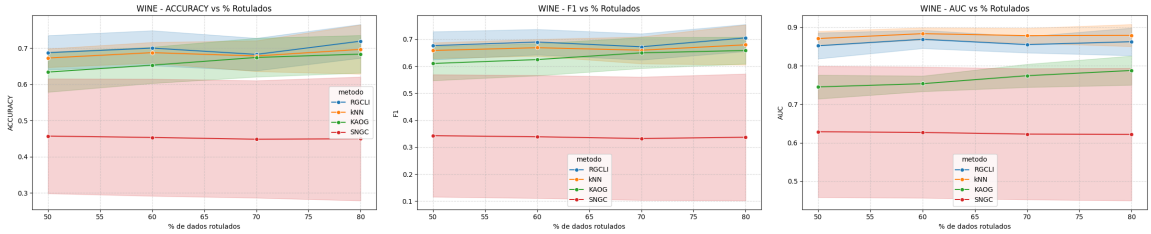


Figura 10 – Desempenho dos métodos na base **Wine** em função da porcentagem de dados rotulados. As curvas mostram média e desvio padrão ao longo das repetições.

Base Breast Cancer

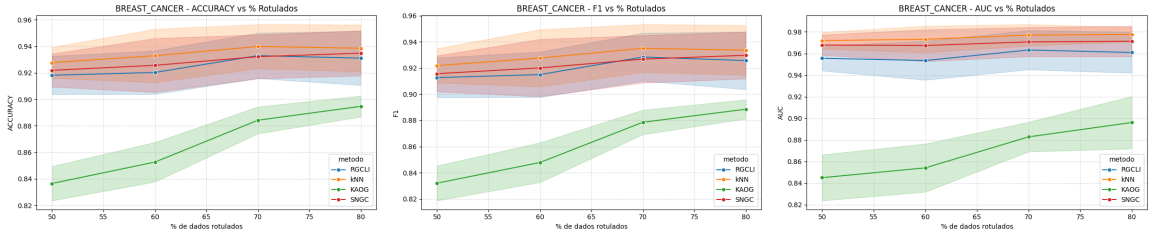


Figura 11 – Desempenho dos métodos na base **Breast Cancer** em função da porcentagem de dados rotulados. As curvas mostram média e desvio padrão ao longo das repetições.

Base Diabetes

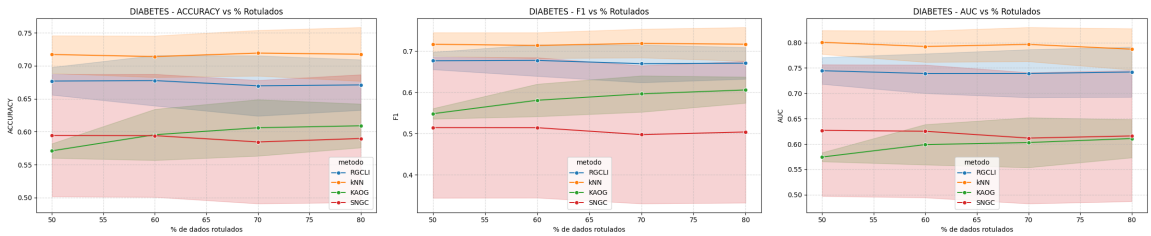


Figura 12 – Desempenho dos métodos na base **Diabetes** em função da porcentagem de dados rotulados. As curvas mostram média e desvio padrão ao longo das repetições.

Base 20 Newsgroups

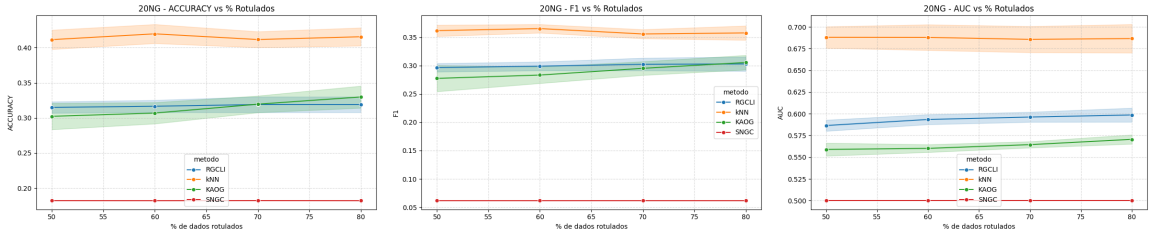


Figura 13 – Desempenho dos métodos na base **20 Newsgroups** em função da porcentagem de dados rotulados. As curvas mostram média e desvio padrão ao longo das repetições.

4.3 Variação dos Parâmetros

4.3.0.1 RGCLI.

O desempenho variou de forma significativa com o parâmetro k_i, sendo observado um ponto de máximo desempenho intermediário entre os extremos testados (2 a 8). Valores muito baixos de k_i resultam em grafos esparsos e baixa propagação, enquanto valores muito altos reduzem a seletividade das conexões confiáveis. Observou-se melhor desempenho com k_i entre 4 e 6.

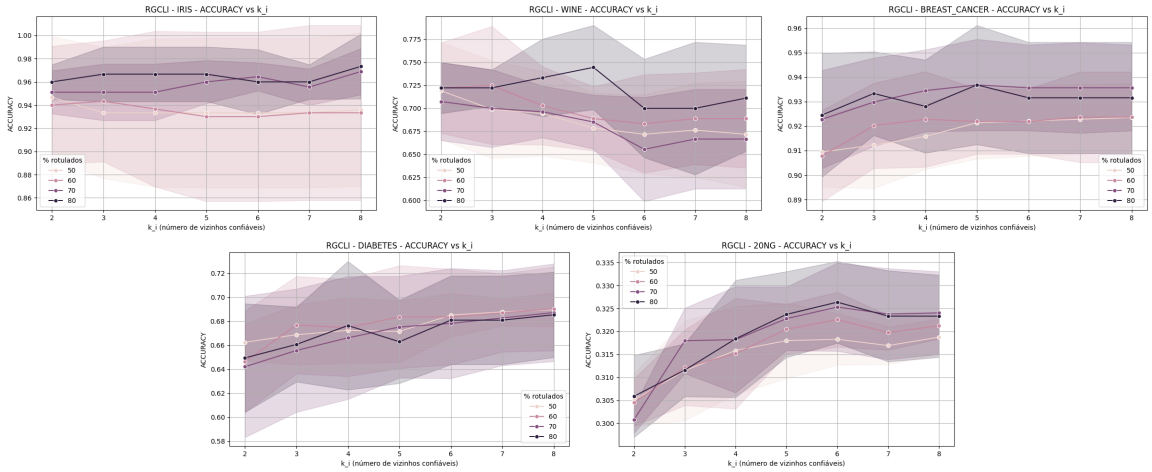


Figura 14 – Acurácia do método RGCLI em diferentes bases de dados, variando o parâmetro k_i.

Observa-se que o parâmetro k_i, que controla a seletividade das conexões confiáveis, exerce forte influência no desempenho. Para a maioria dos datasets, valores intermediários de k_i (geralmente entre 4 e 6) tendem a maximizar a acurácia. Valores baixos (ex: $k_i = 2$) resultam em grafos muito esparsos, dificultando a propagação dos rótulos. Por outro lado, valores muito altos tornam as conexões menos discriminativas, reduzindo a eficácia da supervisão.

4.3.0.2 KAOG.

A variação de $K_{\max}$ apresentou baixo impacto nas métricas, indicando que a seleção de componentes puras é mais influenciada pela topologia local do que pela profundidade da busca. A limitação estrutural do método, que evita conexões entre diferentes classes rotuladas, resultou em grafos com baixa conectividade, prejudicando a propagação.

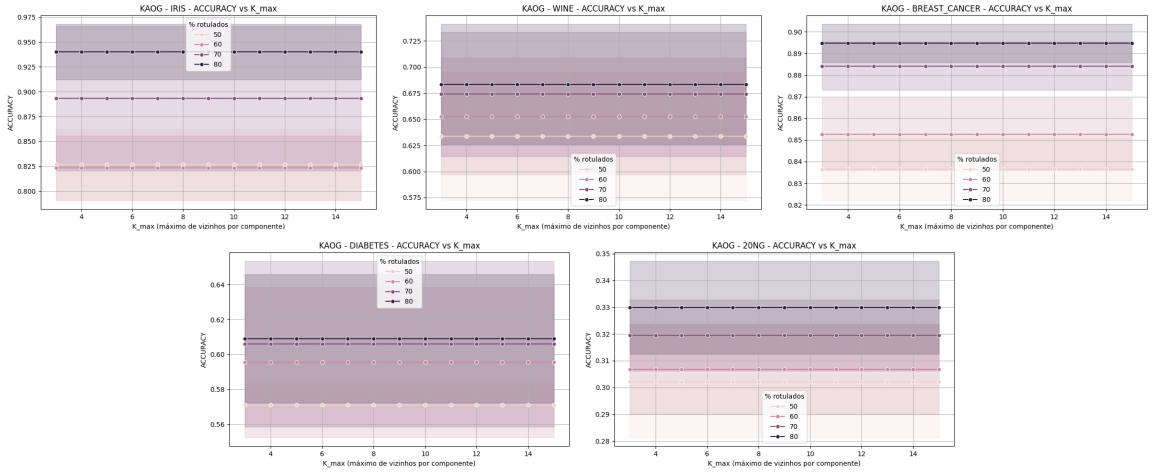


Figura 15 – Acurácia do método KAOG em diferentes bases de dados, variando o parâmetro K_{max}.

Para o KAOG, a variação de K_{max} não resultou em grandes mudanças na acurácia. Isso sugere que a performance do método é menos sensível à profundidade da busca por componentes e mais determinada pela estrutura local imposta pela pureza das classes. A limitação de conectar apenas nós rotulados com o mesmo rótulo pode restringir a conectividade do grafo final, o que se reflete em acurácias mais baixas e menos responsivas ao ajuste de K_{max}.

4.3.0.3 SNGC.

Os parâmetros k (número final de vizinhos) e conf_{min} (limiar de confiança) mostraram impacto considerável: limiares muito altos reduzem drasticamente o número de arestas aceitas, dificultando a propagação. Um valor intermediário (0.7) tende a equilibrar bem a qualidade e quantidade de conexões. O uso de amostragem parcial nos pares de treinamento do SVM foi essencial para reduzir o tempo de execução sem comprometer o desempenho médio.

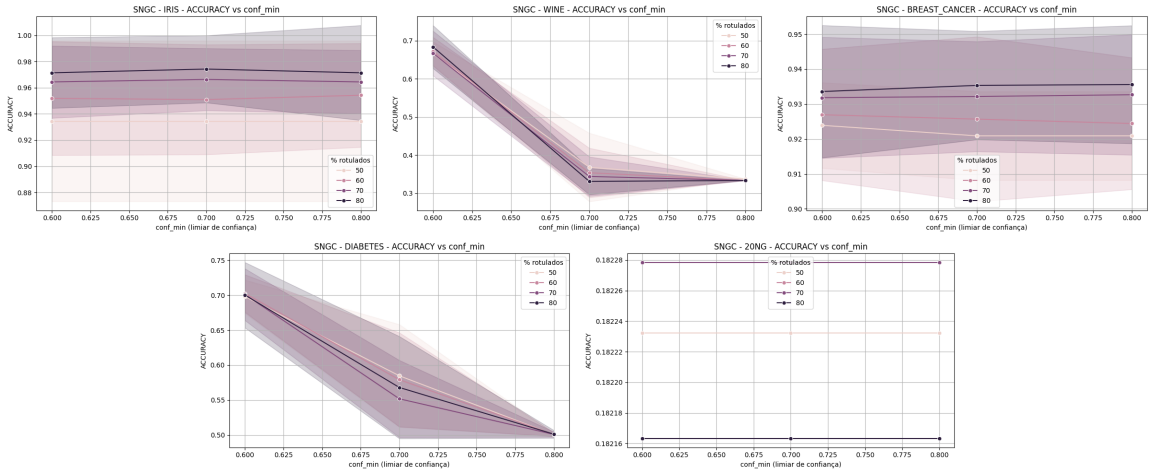


Figura 16 – Acurácia do método SNGC em diferentes bases de dados, variando o limiar de confiança *conf_min*.

A acurácia do SNGC mostra sensibilidade ao limiar de confiança conf_{min}. Limiar muito alto (ex: 0,8) gera grafos excessivamente esparsos, o que prejudica a propagação. Limiar baixo (ex: 0,6) pode permitir muitas conexões ruidosas. Em geral, o valor intermediário de 0,7 ofereceu o melhor equilíbrio entre cobertura e precisão das arestas, especialmente em bases mais complexas como *20ng*.

4.3.0.4 kNN.

O desempenho do método kNN variou de forma moderada com o número de vizinhos k, especialmente em datasets mais complexos. Valores muito baixos de k tendem a gerar grafos mais esparsos e menos robustos à variação dos dados, enquanto valores mais altos promovem maior conectividade, favorecendo a propagação de rótulos. Observou-se que valores de k

entre 7 e 11 geralmente proporcionaram melhor equilíbrio entre conectividade e seletividade, resultando em acurácias mais elevadas.

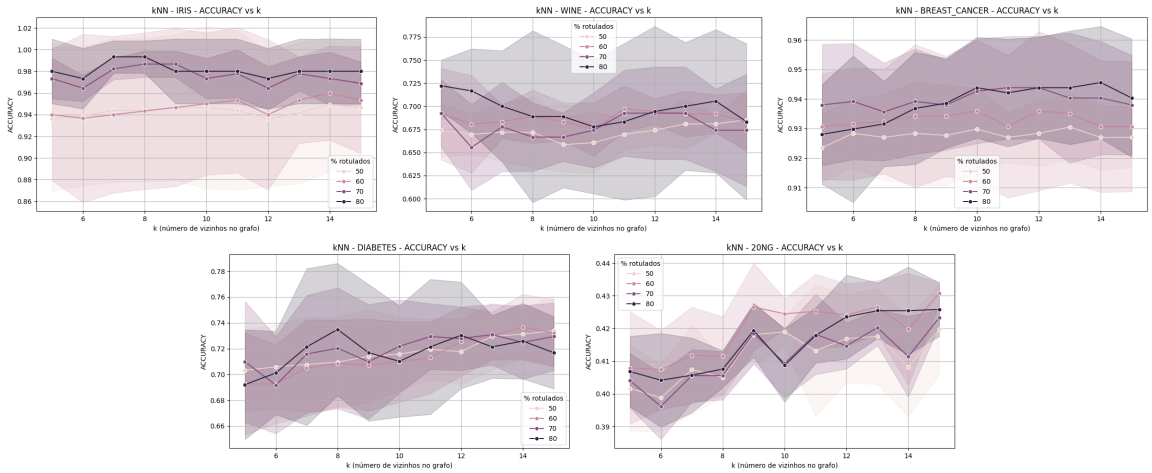


Figura 17 – Acurácia do método kNN em diferentes bases de dados, variando o número de vizinhos k.

O comportamento do kNN é mais estável, com desempenho crescente ou saturado à medida que k aumenta. Como o método não utiliza rótulos na construção do grafo, ele depende fortemente da densidade de vizinhança. Valores muito baixos de k tendem a sub-representar a topologia global, enquanto valores muito altos diluem a separação entre classes. Em muitos casos, $k = 10$ ou próximo disso oferece boa performance.

Nesta seção, avaliamos como os parâmetros internos de cada método influenciam seu desempenho em termos de acurácia, considerando diferentes conjuntos de dados. As curvas apresentadas representam a média da acurácia obtida ao longo das repetições.

4.4 Análise por Dataset

Nos conjuntos `Iris` e `Wine`, com baixa sobreposição entre classes, todos os métodos tiveram bom desempenho. O kNN e o RGCLI destacaram-se com acurácia acima de 0,90.

Já em bases como `Breast Cancer` e `Diabetes`, o desempenho caiu de forma mais homogênea, com destaque ainda para o kNN, que manteve estabilidade. O SNGC obteve resultados intermediários, mas com maior variação.

No caso da `20NG`, os métodos supervisionados apresentaram vantagem, especialmente o SNGC, que se beneficiou da representação TF-IDF de alta dimensionalidade para capturar relações semânticas. O KAOG teve dificuldades nesse cenário, reforçando sua limitação para estruturas mais complexas.

4.5 Resumo Global dos Resultados

Tabela 2 – Média e desvio padrão das métricas globais por método, considerando todos os datasets e níveis de rotulagem.

Método	Acurácia (média ± desvio)	F1-score (média ± desvio)	AUC (média ± desvio)
kNN	0.742 ± 0.222	0.727 ± 0.242	0.866 ± 0.128
RGCLI	0.712 ± 0.254	0.705 ± 0.260	0.827 ± 0.162
SNGC	0.622 ± 0.327	0.557 ± 0.384	0.742 ± 0.225
KAOG	0.662 ± 0.230	0.648 ± 0.238	0.744 ± 0.160

A Tabela 2 apresenta o resumo quantitativo geral de desempenho médio e variabilidade (desvio padrão) de cada método de construção de grafos, considerando todos os experimentos realizados. Os dados reforçam os resultados discutidos anteriormente e servem como base para a comparação estatística subsequente.

4.6 Análise Estatística

Para investigar se as diferenças observadas nas métricas de desempenho entre os métodos de construção de grafos são estatisticamente significativas, foi utilizada a biblioteca `autorank`, que automatiza a condução dos testes estatísticos apropriados com base nas propriedades dos dados.

Pré-testes de normalidade e homogeneidade

Inicialmente, foram realizados testes de normalidade (Shapiro-Wilk) e homogeneidade de variância (Bartlett). Os resultados não rejeitaram a hipótese nula de normalidade para nenhum dos métodos em nenhuma das métricas (p-valor mínimo observado: 0,172), tampouco rejeitaram a homocedasticidade (homogeneidade de variância). Isso valida a aplicação do teste paramétrico **ANOVA de medidas repetidas**.

Resultados do ANOVA e pós-teste de Tukey HSD

O ANOVA detectou diferenças estatisticamente significativas entre os métodos nas três métricas analisadas:

- **Acurácia**: $p = 0{,}038$
- **F1-score**: $p = 0{,}029$
- **AUC**: $p = 0{,}010$

Esses valores indicam que, para cada métrica, ao menos um dos métodos apresentou desempenho significativamente diferente dos demais.

O teste **post hoc de Tukey HSD** revelou as comparações com significância estatística. A seguir, apresentamos uma análise qualitativa dos métodos com base nos resultados:

- **kNN (baseline)**: Apresentou os melhores resultados médios em todas as métricas (Acurácia: 0,742, F1: 0,727, AUC: 0,866), sendo estatisticamente superior a métodos como KAOG e SNGC. Apesar de ser um método não supervisionado, sua simplicidade e densidade uniforme explicam sua vantagem em cenários com alta proporção de rótulos. No entanto, tende a saturar em cenários ruidosos ou mais escassos.
- **RGCLI**: Obteve desempenho competitivo, com destaque para a F1-score (0,705) e acurácia (0,712), sendo estatisticamente próximo ao kNN e superior ao KAOG e SNGC. O método demonstrou robustez em diferentes cenários, sugerindo que a seleção de vizinhos confiáveis foi eficaz.
- **SNGC**: Apesar de utilizar informações supervisionadas via SVM, apresentou maior variabilidade nos resultados (ex: AUC: $0{,}742 \pm 0{,}225$), e desempenho inferior ao RGCLI e kNN. A sensibilidade à amostragem e o custo computacional elevado podem ter afetado sua estabilidade.
- **KAOG**: Apresentou os piores desempenhos médios nas três métricas (ex: F1-score: 0,648, AUC: 0,744), sendo consistentemente agrupado no ranking inferior. Sua limitação em conectar apenas amostras rotuladas da mesma classe reduz a conectividade do grafo, prejudicando a propagação. Mesmo com extensões para conectar nós não representados, os resultados permaneceram inferiores.

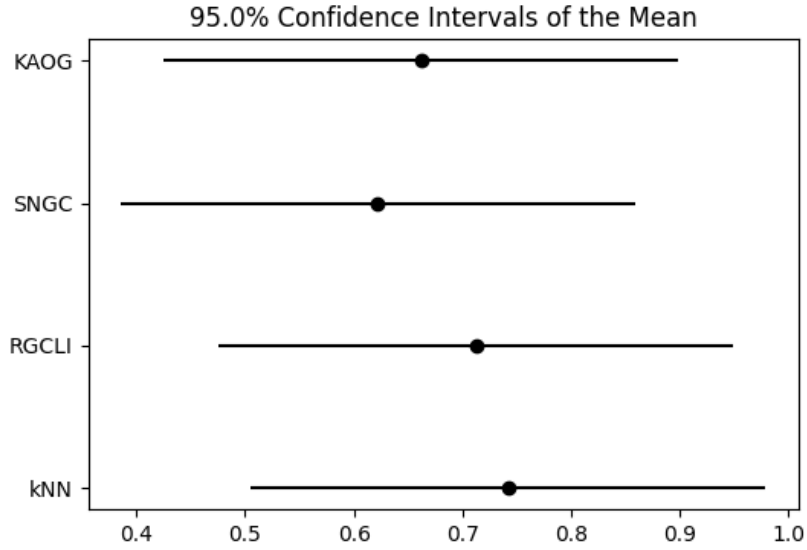


Figura 18 – Ranking médio dos métodos com intervalo de confiança – métrica de Acurácia.

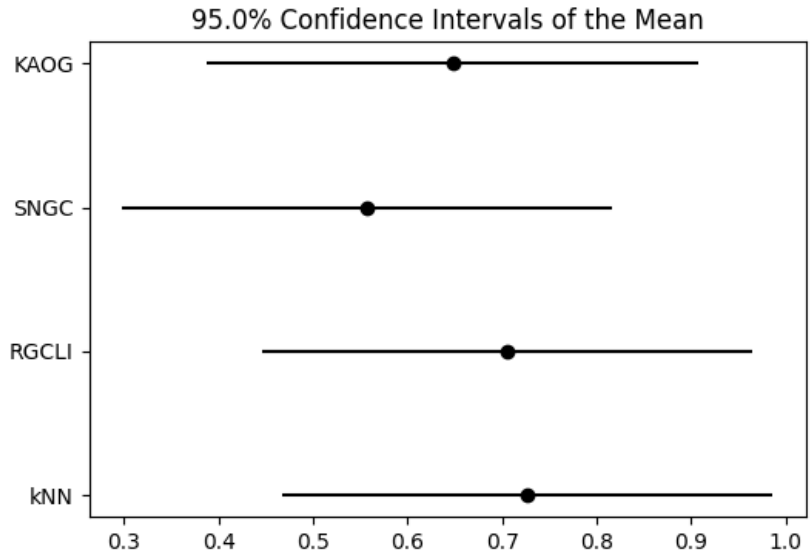


Figura 19 – Ranking médio dos métodos com intervalo de confiança – métrica de F1-score.

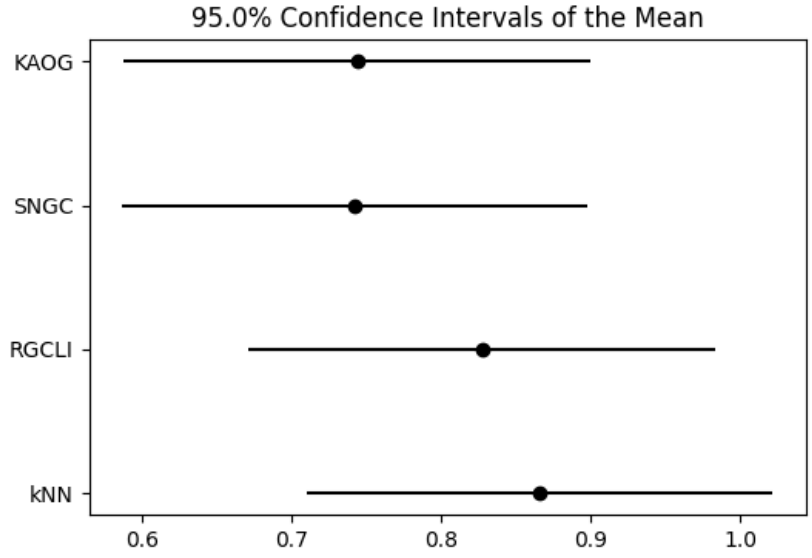


Figura 20 – Ranking médio dos métodos com intervalo de confiança – métrica de AUC.

4.7 Síntese e Limitações

Os resultados estatísticos reforçam que métodos que equilibram supervisão com densidade topológica tendem a apresentar melhor desempenho na tarefa de propagação de rótulos. O RGCLI demonstrou bom desempenho médio e estabilidade, configurando-se como alternativa promissora. O SNGC apresentou potencial especialmente em bases textuais, mas exigiu maior tempo de execução. Já o KAOG mostrou desempenho consistentemente inferior, refletindo limitações práticas de conectividade e propagação eficiente.

Como limitações do trabalho, destacam-se: (i) a não inclusão de técnicas de pré-processamento avançado nos textos da `20NG`, (ii) a ausência de tuning fino para todos os parâmetros de cada método e (iii) a escolha de apenas um algoritmo de propagação (Label Propagation), o que restringe a análise a esse cenário específico.

5 CONCLUSÃO

Este trabalho analisou o impacto da escolha do método de construção de grafos no desempenho de algoritmos de propagação de rótulos em cenários semi-supervisionados. Foram avaliados quatro métodos — **kNN**, **RGCLI**, **KAOG** e **SNGC** — aplicados a cinco conjuntos de dados com diferentes níveis de rotulagem e complexidade.

Os resultados mostram que, de modo geral, o método **kNN (baseline)** apresentou o melhor desempenho médio nas métricas de acurácia, F1-score e AUC, destacando-se por sua simplicidade e conectividade estável mesmo sem utilizar informações supervisionadas na construção do grafo.

Entre os métodos semi-supervisionados, o **RGCLI** demonstrou desempenho competitivo e robusto, sendo estatisticamente próximo ao kNN e superior aos demais métodos supervisionados. O **SNGC**, embora tenha apresentado maior variabilidade nos resultados, destacou-se especialmente na base textual **20NG**, evidenciando sua capacidade de lidar com representações vetoriais esparsas por meio da abordagem supervisionada baseada em SVM. O **KAOG**, por sua vez, obteve os piores resultados médios, o que pode ser atribuído às restrições impostas por sua estrutura, que limita a conectividade entre instâncias de diferentes classes rotuladas.

Além do desempenho dos métodos, observou-se que a estrutura intrínseca dos conjuntos de dados influenciou significativamente os resultados obtidos. Bases como **Iris**, **Wine** e **Breast Cancer**, que possuem separação clara entre classes, baixa sobreposição e número reduzido de instâncias, favoreceram abordagens mais simples como o kNN, cuja conectividade densa garante boa propagação de rótulos mesmo sem orientação supervisionada.

Outro fator importante foi a alta proporção de rótulos considerada em muitos cenários experimentais, o que contribuiu para a formação de grafos bem conectados em todos os métodos. Isso pode ter reduzido o impacto relativo das estratégias supervisionadas na construção dos grafos, uma vez que a densidade de rótulos já era suficiente para garantir uma boa propagação em abordagens não supervisionadas.

Em contrapartida, em bases mais desafiadoras como a **20NG**, que possui alta dimensionalidade, dados textuais e estrutura mais esparsa, métodos supervisionados como o SNGC mostraram vantagens por sua capacidade de selecionar conexões mais relevantes com base em conhecimento prévio.

Do ponto de vista metodológico, este estudo também evidenciou a influência direta da proporção de rótulos disponíveis e da escolha dos parâmetros internos de cada método no

desempenho final da propagação. Parâmetros como k_i (RGCLI), conf_{min} (SNGC) e K_{max} (KAOG) mostraram-se sensíveis ao tipo de dado e à densidade do grafo resultante, sendo essenciais para alcançar bons resultados.

Conclui-se, portanto, que a escolha do método de construção de grafos deve ser feita com base não apenas na quantidade de rótulos disponíveis, mas também levando em consideração as características estruturais dos dados e o objetivo da tarefa. Métodos como o RGCLI demonstram-se especialmente promissores por aliar informação supervisionada à estrutura local dos dados, mantendo boa conectividade e desempenho estável.

6 TRABALHOS FUTUROS

Como próximos passos, sugerem-se as seguintes direções de investigação:

- Avaliação de outros algoritmos de propagação de rótulos além do Label Propagation, como o Label Spreading ou métodos baseados em aprendizado profundo;
- Inclusão de técnicas de pré-processamento textual mais sofisticadas, como embeddings do tipo *word2vec*, *BERT* ou redução de dimensionalidade supervisionada para dados textuais;
- Análise da aplicabilidade dos métodos de construção de grafos em tarefas distintas da classificação, como agrupamento (clustering), detecção de anomalias ou recomendação.

REFERÊNCIAS

BARTLETT, M. S. Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, The Royal Society, v. 160, n. 901, p. 268–282, 1937.

BERTON, L. et al. Rgcli: Robust graph that considers labeled instances for semi-supervised learning. *Neurocomputing*, Elsevier, v. 226, p. 238–248, 2017.

BERTON, L. et al. Rgcli: Robust graph that considers labeled instances for semi-supervised learning. *Neurocomputing*, Elsevier, v. 226, p. 238–248, 2017.

BERTON, L.; LOPES, A. D. A. Graph construction based on labeled instances for semi-supervised learning. In: IEEE. *2014 22nd international conference on pattern recognition*. [S.l.], 2014. p. 2477–2482.

BERTON, L.; LOPES, A. de A. Graph construction for semi-supervised learning. In: *IJCAI*. [S.l.: s.n.], 2015. p. 4343–4344.

BLUM, A.; MITCHELL, T. Combining labeled and unlabeled data with co-training. In: *Proceedings of the eleventh annual conference on Computational learning theory (COLT)*. [S.l.: s.n.], 1998. p. 92–100.

BREVE, F.; ZHAO, L. Particle competition and cooperation in networks for semi-supervised learning with concept drift. In: IEEE. *The 2012 International Joint Conference on Neural Networks (IJCNN)*. [S.l.], 2012. p. 1–6.

BREVE, F. et al. Particle competition and cooperation in networks for semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 24, n. 9, p. 1686–1698, 2011.

CALDAS, W. L. *Co-MLM: Uma abordagem baseada em co-training para aprendizado semi-supervisionado*. Dissertação (Mestrado) — Universidade Federal do Ceará, 2017. Disponível em: <https://repositorio.ufc.br/bitstream/riufc/24913/1/2017_dis_wlcaldas.pdf>.

CARNEIRO, M. G. et al. Network-based data classification: combining k-associated optimal graphs and high-level prediction. *Journal of the Brazilian Computer Society*, Springer, v. 20, p. 1–14, 2014.

CHEN, C. et al. Interactive graph construction for graph-based semi-supervised learning. *IEEE Transactions on Visualization and Computer Graphics*, IEEE, v. 27, n. 9, p. 3701–3716, 2021.

CHEN, Z.; CAO, H.; CHANG, K. C.-C. Graphebm: Energy-based graph construction for semi-supervised learning. In: IEEE. *2020 IEEE International Conference on Data Mining (ICDM)*. [S.l.], 2020. p. 62–71.

DATACAMP. *Aprendizado semi supervisionado + Label Propagation*. 2023. <https://www.youtube.com/watch?v=PD9ainVUk3o>. Acesso em: 27 abr. 2025.

DEMšAR, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, v. 7, n. Jan, p. 1–30, 2006.

DORNAIKA, F.; TRABOULSI, Y. E. Joint sparse graph and flexible embedding for graph-based semi-supervised learning. *Neural Networks*, Elsevier, v. 114, p. 91–95, 2019.

FERRI, C.; HERNáNDEZ-ORTEGA, J.; HERNáNDEZ-ORALLO, J. Comparative evaluation of classification methods using roc curves. In: *Lecture Notes in Artificial Intelligence*. [S.l.]: Springer, 2001. v. 2086, p. 101–112.

FIELD, A. *Discovering statistics using IBM SPSS statistics*. 5. ed. [S.l.]: Sage Publications, 2018.

FONTES, A. F. C. *Métodos baseados em grafos e auto-treinamento para aprendizado semi-supervisionado*. Dissertação (Mestrado) — Instituto Politécnico do Porto, 2023. Disponível em: <https://recipp.ipp.pt/bitstream/10400.22/23869/1/DM_AndreFontes_2023_MEI.pdf>.

GHAHRAMANI, Z. Unsupervised learning. In: *Summer school on machine learning*. [S.l.]: Springer, 2003. p. 72–112.

GRAFOS: conceitos, algoritmos e aplicações. Elsevier, 2012. ISBN 9788535257182. Disponível em: <https://books.google.com.br/books?id=cyrtyQEACAAJ>.

GUERREIRO, L. M. *Aprendizado semi-supervisionado utilizando estruturas de comunidades em grafos*. Tese (Doutorado) — Universidade Estadual Paulista, 2019. Disponível em: <https://repositorio.unesp.br/bitstream/11449/151923/3/guerreiro_l_me_sjrp.pdf>.

HERBOLD, S. Autorank: a python package for automated ranking of classifiers. *arXiv preprint arXiv:2006.10108*, 2020.

IBM. *Pseudo-labeling explained*. 2023. Disponível em: <https://www.ibm.com/topics/pseudo-labeling>.

KANG, M. et al. Autoencoder-based graph construction for semi-supervised learning. In: SPRINGER. *European conference on computer vision*. [S.l.], 2020. p. 500–517.

KOTSIANTIS, S. B. et al. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, Amsterdam, v. 160, n. 1, p. 3–24, 2007.

LOPES, A. A. et al. Classification based on the optimal k-associated network. In: SPRINGER. *Complex Sciences: First International Conference, Complex 2009, Shanghai, China, February 23-25, 2009. Revised Papers, Part 1 1*. [S.l.], 2009. p. 1167–1177.

MAIER, M.; HEIN, M.; LUXBURG, U. V. Cluster identification in nearest-neighbor graphs. In: SPRINGER. *Algorithmic Learning Theory: 18th International Conference, ALT 2007, Sendai, Japan, October 1-4, 2007. Proceedings 18*. [S.l.], 2007. p. 196–210.

MONARD, M. C. et al. Uma introdução ao aprendizado simbólico de máquina por exemplos. 1997.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

RAMOS, L. R. *Técnicas de propagação de rótulos semi-supervisionadas com aplicação em classificação de dados*. Dissertação (Mestrado) — Instituto Federal do Espírito Santo, 2020. Disponível em: <https://repositorio.ifes.edu.br/bitstream/handle/123456789/1826/TeseMestrado_Leandro%20Rodrigues%20Ramos.pdf?sequence=1&isAllowed=y>.

RAMOS, L. R. et al. Geração semiautomática de valores de referência para identificação de obstruções em lingotamento contínuo. In: SBC. *Seminário Integrado de Software e Hardware (SEMISH)*. [S.l.], 2020. p. 116–127.

REZENDE, S. O. *Sistemas Inteligentes: Fundamentos e Aplicações*. Barueri, SP: Editora Manole Ltda, 2003. ISBN 8520416837.

ROHBAN, M. H.; RABIEE, H. R. Supervised neighborhood graph construction for semi-supervised classification. *Pattern Recognition*, Elsevier, v. 45, n. 4, p. 1363–1372, 2012.

SANTOS, G. Giordani dos. *Uma análise sobre a acurácia e a escalabilidade de algoritmos de detecção de comunidades*. Tese (Doutorado) — PUCRS, 2023. Disponível em: <https://tede2.pucrs.br/tede2/bitstream/tede/10560/2/GABRIEL_GIORDANI_DOS_SANTOS_DIS.pdf>.

SANTOS, M. S. *Algoritmos de aprendizado semi-supervisionado baseados em grafos e suas aplicações em dados biológicos*. Dissertação (Mestrado) — Universidade Estadual Paulista, 2014. Disponível em: <https://bdtd.ibict.br/vufind/Record/UNSP_05205cd74709d120789744ff9d13eb69>.

SANTOS, M. S. *Algoritmos de aprendizado semi-supervisionado baseados em grafos e suas aplicações em dados biológicos*. Dissertação (Mestrado) — Universidade Estadual Paulista, 2016. Disponível em: <https://repositorio.unesp.br/items/8653d212-2fb4-44ef-9f07-b21f1439f318>.

scikit-learn developers. *LabelSpreading - scikit-learn 1.6.1 documentation*. 2024. <https://scikit-learn.org/stable/modules/generated/sklearn.semi_supervised.LabelSpreading.html>. Acesso em: 27 abr. 2025.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). *Biometrika*, Oxford University Press, v. 52, n. 3/4, p. 591–611, 1965.

TAHERKHANI, F.; KAZEMI, H.; NASRABADI, N. M. Matrix completion for graph-based deep semi-supervised learning. In: *Proceedings of the AAAI conference on artificial intelligence*. [S.l.: s.n.], 2019. v. 33, n. 01, p. 5058–5065.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introdução ao Data Mining*. [S.l.]: Pearson, 2009.

TUKEY, J. W. *The problem of multiple comparisons*. 1953. Unpublished manuscript, Princeton University.

VEGA-OLIVEROS, D. A. et al. Regular graph construction for semi-supervised learning. In: IOP PUBLISHING. *Journal of physics: Conference series*. [S.l.], 2014. v. 490, n. 1, p. 012022.

ZHOU, Z.-H.; LI, M. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on Knowledge and Data Engineering*, v. 17, n. 11, p. 1529–1541, 2005.

ZHU, X. *Semi-supervised learning literature survey*. [S.l.], 2005. Computer Sciences Technical Report.

ZHUANG, L. et al. Label information guided graph construction for semi-supervised learning. *IEEE Transactions on Image Processing*, IEEE, v. 26, n. 9, p. 4182–4192, 2017.