A Robust Pseudo-label Reevaluation Strategy for the Self-training Algorithm

Luiz M. S. Silva[1], **Renan M. R. A. Costa**[1], **José A. A. Paiva**[1]
Arthur C. Gorgônio[2], **Karliane M. O. Vale**[3], **Flavius L. Gorgônio**[3]

[1]Laboratório de Inteligência Computacional Aplicada a Negócios (LABICAN)
Universidade Federal do Rio Grande do Norte (UFRN)
Rua Joaquim Gregório, 296 – 59.300-000 – Caicó – RN – Brasil

[2]Departamento de Informática e Matemática Aplicada (DIMAP)
Universidade Federal do Rio Grande do Norte (UFRN)
Campus Universitário – Lagoa Nova – 59.078-970 – Natal – RN – Brasil

[3]Departamento de Computação e Tecnologia (DCT)
Universidade Federal do Rio Grande do Norte (UFRN)
Rua Joaquim Gregório, 296 – 59.300-000 – Caicó – RN – Brasil

{luiz.santos.090, renan.costa.117, jose.alves.092}@ufrn.edu.br
{arthur.gorgonio.099}@ufrn.edu.br
{karliane.vale, flavius.gorgonio}@ufrn.br

Abstract. *This paper proposes an extension of the self-training algorithm with iterative pseudo-label reevaluation, using the silhouette metric to identify and remove noisy instances, and ensembles with weighted voting to support decisions in low-confidence scenarios. The approach aims to mitigate error propagation and enhance the robustness of semi-supervised learning. Evaluations conducted on 18 datasets demonstrated superior performance compared to the original self-training algorithm in terms of accuracy, F1-score, and stability, especially in scenarios with limited labeled data.*

Resumo. *Este trabalho propõe uma extensão do algoritmo Self-Training com reavaliação iterativa de pseudo-rótulos, utilizando a métrica de silhueta para identificar e remover instâncias ruidosas, e um comitê de classificadores com votação ponderada para reforçar decisões em casos de baixa confiança. A abordagem visa mitigar a propagação de erros e aumentar a robustez do aprendizado semissupervisionado. Avaliações realizadas em 18 bases de dados demonstraram desempenho superior ao self-training original em termos de acurácia, F1-score e estabilidade, especialmente em cenários com poucos dados rotulados.*

1. Introdução

O aumento exponencial de dados não rotulados em diversas áreas de aplicação, como saúde, finanças e visão computacional, tem motivado o desenvolvimento de técnicas de aprendizado semissupervisionado (SSL), que são capazes de reduzir a dependência de grandes quantidades de dados rotulados manualmente [Gomes et al. 2021]. Dentre as abordagens de SSL, o método self-training se destaca pela sua simplicidade e eficácia: inicialmente, um classificador é treinado com um pequeno conjunto de dados rotulados e,

de forma progressiva, instâncias não rotuladas recebem pseudo-rótulos (rótulos preditos pelo classificador treinado) que são incorporados ao conjunto de treinamento, com base na confiança atribuída pelo modelo [Amini et al. 2025]. Contudo, a propagação de erros de pseudo-rotulagem pode degradar a qualidade do conjunto de treinamento, especialmente em cenários de alta complexidade ou dados ruidosos [Xie et al. 2020, Sun et al. 2021].

Em métodos de pseudo-rotulagem, como o *self-training*, assume-se que os exemplos classificados com maior confiança são também os mais corretos. Entretanto, essa premissa nem sempre se sustenta, principalmente quando há sobreposição de classes ou ruído nos dados [Wang et al. 2023]. Rótulos incorretos podem ser incorporados ao conjunto de treinamento e reforçados nas iterações subsequentes, levando à propagação de erros e à degradação do desempenho do classificador [Sun et al. 2021]. Este efeito é particularmente problemático em conjuntos de dados desbalanceados, onde classes minoritárias tendem a ser sub-representadas nos pseudo-rótulos gerados automaticamente [Xie et al. 2020].

Apesar desses desafios, há uma lacuna na literatura no que diz respeito a mecanismos de reavaliação periódica dos pseudo-rótulos atribuídos. A maioria dos métodos foca na seleção inicial de instâncias confiáveis, mas poucos abordam explicitamente a remoção ou correção de pseudo-rótulos que, ao longo do processo iterativo, possam ter se tornado inconsistentes com a estrutura dos dados [Amini et al. 2025]. Abordagens como o Noisy Student introduzem estratégias de suavização e aumento de dados, mas não realizam uma reavaliação estruturada com base na qualidade dos clusters formados [Xie et al. 2020]. Por outro lado, em [Vale et al. 2021] foi proposto o uso de um limiar ajustável a cada iteração com o objetivo de minimizar a aleatoriedade com que as instâncias são escolhidas no processo de rotulagem.

Portanto, embora existam técnicas de reamostragem e seleção de amostras para lidar com rótulos potencialmente errôneos em SSL, ainda há uma carência de critérios sistemáticos e robustos para identificar e descartar esses rótulos de forma eficiente, o que pode comprometer a qualidade do modelo final [Gomes et al. 2021]. Diante do exposto, identificou-se uma oportunidade para aprimorar o processo de pseudo-rotulagem no algoritmo *self-training* por meio da introdução de um mecanismo de reavaliação periódica dos pseudo-rótulos, baseado na métrica de silhueta, que mede a coesão e separação dos clusters formados. Além disso, propomos a utilização de um comitê de classificadores que, por meio de votação ponderada, decide a atribuição de novos pseudo-rótulos às instâncias com baixa qualidade de cluster, agregando a diversidade e robustez ao processo de rotulagem.

Diante das limitações nas abordagens tradicionais, este trabalho apresenta três principais contribuições: (i) o mecanismo de reavaliação de pseudo-rótulos baseado na métrica de silhueta, capaz de identificar e remover rótulos potencialmente incorretos durante o processo iterativo de autoaprendizagem; (ii) o comitê de classificadores com votação ponderada, que amplia a robustez do processo de rotulagem quando o classificador especialista apresenta baixa confiança; e (iii) a avaliação experimental abrangente em 18 bases de dados de diferentes domínios, evidenciando ganhos consistentes em desempenho e estabilidade frente ao self-training tradicional. Até onde sabemos, esta é a primeira proposta a integrar esses dois mecanismos de forma sinérgica no contexto de SSL.

Este artigo está organizado da seguinte forma: a Seção 2 apresenta os principais conceitos sobre aprendizado semissupervisionado, Self-Training, reavaliação de rótulos

e comitê de classificadores. A Seção 3 revisa os principais trabalhos relacionados a esta pesquisa. A Seção 4 descreve em detalhes a metodologia experimental, incluindo as bases utilizadas e o desenho dos experimentos. A Seção 5 apresenta os resultados e discussão dos experimentos. Por fim, a Seção 6 conclui o estudo e sugere trabalhos futuros.

2. Aspectos Teóricos

2.1. Aprendizado Semissupervisionado

O SSL utiliza um conjunto de dados rotulados D_l, pequeno, e um maior com dados não rotulados D_u, explorando a estrutura e a distribuição intrínseca dos dados não rotulados para ajustar melhor as fronteiras de decisão e melhorar a capacidade de generalização dos modelos de aprendizado [Zhu and Goldberg 2009]. O SSL surgiu como uma solução para cenários reais, nos quais a rotulagem manual é onerosa ou inviável, como em aplicações na área médica, bioinformática e processamento de linguagem natural [Amini et al. 2025].

O aprendizado semissupervisionado tende a ser mais eficaz quando o conjunto de dados rotulados (D_l) representa bem a distribuição subjacente do conjunto não rotulado (D_u), isto é, quando ambos os conjuntos seguem a mesma distribuição estatística. Nessa condição, os algoritmos conseguem explorar padrões estruturais nos dados não rotulados para melhorar a definição das fronteiras de decisão, mesmo dispondo de uma quantidade limitada de exemplos anotados [Chapelle et al. 2006]. Entretanto, um dos desafios do SSL é a incorreta atribuição de pseudo-rótulos, que pode ocorrer devido à sobreposição entre classes ou à presença de ruído nos dados, levando à propagação de erros ao longo das iterações. Além disso, a presença de classes desbalanceadas dificulta o aprendizado, uma vez que as classes minoritárias tendem a ser sub-representadas tanto nos dados rotulados quanto na geração de pseudo-rótulos [Wang et al. 2023]. Outro desafio relevante é a alta dimensionalidade dos dados, que pode dificultar a estimação precisa das fronteiras de decisão e afetar negativamente o desempenho dos algoritmos de SSL [Fang et al. 2023].

Entre os diversos algoritmos de SSL, o *self-training* destaca-se por sua simplicidade e aplicabilidade em diversos domínios, tais como, visão computacional, detecção de objetos em imagens, data stream, reconhecimento facial entre outros [Wang et al. 2023, Oymak and Gulcu 2020, Fang et al. 2023, He et al. 2023, Gomes et al. 2021]. O método consiste em treinar inicialmente um classificador f com um pequeno conjunto rotulado D_l e, iterativamente, utilizá-lo para prever rótulos das instâncias não rotuladas D_u. As predições com confiança superior a um limiar t são incorporadas ao conjunto rotulado e o modelo é retreinado com esse novo conjunto. O processo se repete até não restarem instâncias elegíveis para classificação ou atinja algum outro critério de parada [Chapelle et al. 2006].

As principais vantagens do *self-training* incluem sua simplicidade e a facilidade de integração com diversos classificadores. Entretanto, uma vez que uma instância seja classificada pelo algoritmo, aquele rótulo à ela atribuído passa a ser considerado como correto daquele ponto em diante e será utilizado no treinamento do algoritmo nas iterações seguintes. Dessa forma, a técnica é suscetível à propagação de erros: predições incorretas feitas com alta confiança podem ser incorporadas ao conjunto de treinamento, degradando a performance do modelo ao longo das iterações [Sun et al. 2021]. Para mitigar tais limitações, diversas extensões foram propostas. Entre elas o Noisy Student se destaca ao combinar o *self-training* com técnicas de aumento de dados, alcançando excelentes resultados em tarefas de visão computacional [Xie et al. 2020]. Outras variantes opõem

ajustes dinâmicos no limiar de confiança [Li et al. 2019] ou utilizam critérios baseados na entropia e margem das predições [Li et al. 2021].

Nesse contexto, técnicas de SSL continuam a ser uma área ativa de pesquisa, com o objetivo de desenvolver métodos que sejam robustos a ruídos, desbalanceamento e à distribuição complexa dos dados, ampliando suas aplicações práticas. Dessa forma, a proposta apresentada neste trabalho busca mitigar especialmente o desafio da propagação de erros, além da adoção de um comitê de classificadores que promove uma decisão mais robusta na atribuição de pseudo-rótulos, atenuando parcialmente os efeitos negativos do desbalanceamento de classes.

2.2. Reavaliação de Rótulos

A reavaliação de rótulos no contexto de SSL refere-se à prática de revisar periodicamente os pseudo-rótulos atribuídos automaticamente pelo modelo durante o processo iterativo de autoaprendizagem. O objetivo central dessa reavaliação é identificar e corrigir possíveis erros de rotulagem que possam comprometer a qualidade do modelo e induzir à propagação de erros ao longo das iterações [He et al. 2023, Sun et al. 2021].

Conforme explicado anteriormente, os métodos clássicos de self-training baseiam-se na premissa de que as previsões realizadas com maior confiança pelo classificador tendem a estar corretas. No entanto, essa hipótese nem sempre se sustenta, especialmente em contextos com classes sobrepostas ou dados ruidosos, o que pode resultar na propagação de rótulos incorretos para o conjunto de treinamento e, consequentemente, comprometer o desempenho do modelo [Amini et al. 2025]. Por essa razão, surgem abordagens que defendem a necessidade de mecanismos de reavaliação periódica, capazes de filtrar ou ajustar pseudo-rótulos que se revelem inconsistentes ao longo do processo. Tais mecanismos visam diminuir a propagação de erros e aprimorar a robustez e a generalização do modelo [Slack et al. 2020].

Uma das estratégias mais promissoras para conduzir essa reavaliação é o uso de métricas de qualidade de agrupamento (clustering), as quais avaliam a coerência das instâncias pseudo-rotuladas em relação à estrutura dos dados [Fang et al. 2023]. Dentre essas métricas, destaca-se a métrica de silhueta (silhouette), amplamente utilizada para medir a coesão e separação entre clusters, proposta por Rousseeuw [Rousseeuw 1987].

A métrica de silhueta para uma instância i é calculada como:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{1}$$

onde:

$a(i)$ denota a distância média de i para todas as outras instâncias do mesmo cluster;

$b(i)$ corresponde à menor distância média de i para o cluster vizinho.

Valores de silhueta próximos a 1 indicam que a instância está bem agrupada, enquanto valores próximos a –1 sugerem que a instância foi mal alocada ao seu cluster. Em contextos de pseudo-rotulagem, instâncias com silhuetas negativas podem ser interpretadas como prováveis rótulos incorretos [Dinh et al. 2025, Guérin et al. 2024, Xie et al. 2021].

Embora o uso de métricas de agrupamento seja tradicionalmente associado ao aprendizado não supervisionado, sua aplicação no contexto semissupervisionado ainda é relativamente recente e pouco explorada, representando uma oportunidade promissora para melhorar a qualidade do processo de autorrotulagem. Algumas abordagens recentes incorporam mecanismos de filtragem ou suavização dos pseudo-rótulos, como o Noisy Student [Xie et al. 2020] e técnicas baseadas em incerteza ou entropia [Li et al. 2021], mas poucas realizam uma reavaliação estruturada com base em métricas de cluster. Dessa forma, a reavaliação sistemática dos pseudo-rótulos, ancorada em métricas como a silhueta, constitui uma linha de pesquisa relevante, capaz de contribuir para o desenvolvimento de algoritmos semissupervisionados mais robustos, menos suscetíveis à propagação de erros e adaptáveis a cenários com ruído e complexidade estrutural elevada [Liu et al. 2024].

2.3. Comitê de Classificadores

O uso de um comitê de classificadores (*ensemble*) é uma técnica amplamente empregada no aprendizado de máquina para melhorar a robustez e a acurácia das predições por meio da combinação de múltiplos modelos [Li et al. 2022]. Portanto, o comitê desempenha um papel fundamental, especialmente em situações nas quais o classificador especialista não consegue atribuir pseudo-rótulos com confiança elevada. Nesses casos, a diversidade de perspectivas proporcionada pelo comitê permite uma tomada de decisão mais robusta e menos suscetível a erros individuais [Wang et al. 2023].

Nesse contexto, a diversidade entre os modelos que compõem o comitê torna-se um fator crítico para o sucesso da abordagem. Para promover essa diversidade, diferentes algoritmos de SSL podem ser combinados estrategicamente, o que contribui para ampliar a capacidade de generalização do sistema e mitigar vieses individuais dos classificadores [Wang et al. 2023]. Para tanto, o comitê proposto nesta pesquisa foi composto pelos seguintes classificadores: **K-Nearest Neighbors (KNN)**: classificador baseado em instâncias, que atribui o rótulo com base nos k vizinhos mais próximos no espaço de atributos; **Naive Bayes (NB)**: classificador probabilístico fundamentado no teorema de Bayes, que assume independência condicional entre os atributos para calcular a probabilidade de cada classe; **Decision Tree (DT)**: modelo hierárquico e interpretável que realiza divisões recursivas nos atributos para construir regras de decisão com base em medidas de impureza; **Random Forest (RF)**: conjunto de árvores de decisão treinadas por meio da técnica de *bagging*, que promove diversidade entre os modelos e reduz o risco de sobreajuste; **XGBoost**: algoritmo baseado em *boosting*, que combina sequencialmente classificadores fracos para formar um modelo robusto, otimizando uma função de perda com regularização; **Regressão Logística (LR)**: modelo linear amplamente utilizado em tarefas de classificação binária, que estima a probabilidade de ocorrência de uma classe com base em uma função logística; e **Rede Neural Artificial (ANN)**: modelo composto por camadas de neurônios artificiais, capaz de aprender representações complexas e não lineares a partir dos dados [Chapelle et al. 2006, Amini et al. 2025] .

Outro fator importante para o bom desempenho do comitê são as estratégias de combinação, dentre as mais comuns destacam-se: *bagging* que combina vários modelos treinados em subconjuntos diferentes dos dados para reduzir a variância; *boosting* que constrói modelos sequencialmente, cada um corrigindo os erros do anterior para reduzir o viés; e votação que agrega as previsões de múltiplos modelos para decidir o resultado final com base na maioria ou média das respostas [Li et al. 2022]. Neste trabalho, adotamos

um esquema de votação ponderada que confere maior influência aos modelos com melhor desempenho no conjunto rotulado, preservando o equilíbrio entre especialização e diversidade na tomada de decisão do comitê. Dessa forma, cada classificador (f) do comitê recebe um peso (w_f), a saber:

$$w_f = \frac{\text{Acurácia}(f)}{\sum_{i=1}^{n} \text{Acurácia}(f_i)} \tag{2}$$

onde:

n é a quantidade de classificadores do comitê;

$f(i)$ é o i-ésimo classificador do comitê.

3. Trabalhos Relacionados

Abordagens baseadas em *self-training* têm sido propostas para o SSL, destacando-se pela simplicidade e efetividade em diferentes domínios. Na filtragem de pseudo-rótulos, [He et al. 2023] propuseram um modelo probabilístico para identificar e corrigir rótulos ruidosos, enquanto [Slack et al. 2020,Xie et al. 2020], propõem variações do Noisy Student, com ruído sintético e comitês de classificadores para maior robustez. Embora eficientes, essas técnicas normalmente se apoiam em medidas de incerteza, como entropia, sem explorar diretamente a estrutura dos dados via métricas de agrupamento, como a silhueta. Outros trabalhos analisados abordam técnicas de self-training como estratégia central em SSL. [Li et al. 2019, Li et al. 2021] exploram autorrotulagem com políticas aprendidas por meta-aprendizado e seleção consciente baseada em confiança, respectivamente. [Fang et al. 2023] combina self-training com aprendizado auto-supervisionado contrastivo, enquanto [Oymak and Gulcu 2020] oferecem fundamentos teóricos sobre o comportamento estatístico do self-training em larga escala. [Amini et al. 2025], por sua vez, apresentam uma revisão abrangente e atualizada sobre os principais métodos de self-training.

Quanto ao uso de comitês de classificadores em aprendizado semissupervisionado ou múltiplas fontes de decisão, [Radosavovic et al. 2021] propuseram comitês para gerar pseudo-rótulos mais robustos, mas não incorporaram a votação ponderada e nem utilizaram outras métricas para reavaliação dos rótulos. [Wang et al. 2023] propõem um esquema evidencial de geração de pseudo-rótulos, e [Li et al. 2022] utilizam comitês inteligentes para avaliação de risco geotécnico em ambientes instáveis. A seleção ativa e criteriosa de instâncias não rotuladas é proposto por [Vale et al. 2021], com métodos eficientes para seleção de instâncias usando algoritmos de SSL, e em [He et al. 2023], com um mecanismo de correção de pseudo-rótulos voltado à detecção semissupervisionada de objetos.

Por fim, enquanto algumas pesquisas exploram métricas de clustering em aprendizado não supervisionado, seu uso em SSL, combinado ao *self-training* e comitês, ainda não foi explorado. Assim, esta proposta distingue-se por unir a reavaliação de pseudo-rótulos baseada na métrica de clusterização silhouette com um comitê de classificadores com votação ponderada, compondo uma abordagem original e ainda não explorada na literatura.

4. Metodologia

A metodologia proposta baseia-se no processo iterativo de autotreinamento do *self-training* introduzindo um mecanismo de reavaliação de pseudo-rótulos. O Algoritmo 1 descreve o

passo-a-passo desta reavaliação. As entradas para esse algoritmo são: i) dados rotulados (D_l); ii) dados não rotulados (D_u); iii) comitê de classificadores a ser utilizados; iv) limiar de classificação; v) número máximo de iterações; e vi) limiar do *silhouette* para reavaliar as instâncias. Inicialmente, o comitê é criado com todos os classificadores da *pool* (linhas 1-4). Em seguida, os pesos dos classificadores do comitê são ajustados utilizando a Equação 2 (linha 5). Em seguida, o classificador com maior acurácia sobre o conjunto rotulado é selecionado como especialista e se mantém fixo até o final da rotulagem (linha 6).

Algorithm 1: Self-Training com Reavaliação de Rótulos

Entrada : D_l: Dados rotulados
D_u: Dados não rotulados
comite: comitê de classificadores
limiar: limiar para inclusão de instâncias
max_iter: quantidade máxima de iterações
silhouette_limiar: limiar do silhouette para reavaliar as instâncias
Saída : dados rotulados

```
1  for cl in comite do
2  |  cl.treinar_classificador(D_l)
3  end
4  comite.atualizar_pesos(D_l)                    // Utilizar Eq. 2
5  especialista ← argmax(comite)
6  iter ← 0
7  D_{l_original} ← D_l
8  while iter < max_iter do
9  |  iter ← iter + 1
10 |  especialista.treina(D_l)
11 |  instancias_selecionadas ← classifica(especialista, D_u) ≥ limiar
12 |  if len(instancias_selecionadas) = 0 then
13 |  |  D_especialista ← D_l − D_{l_original}      // Instâncias rotuladas
   |  |   pelo especialista
14 |  |  silhouettes ← calcular_silhouettes(D_especialista)
15 |  |  instancias_reavaliar ← seleciona_silhouette(silhouettes,
   |  |   silhouette_limiar)
16 |  |  novos_rotulos ← comite.classifica(instancias_reavaliar)
17 |  |  D_l.atualiza_rotulos(instancias_reavaliar, novos_rotulos)
18 |  else
19 |  |  D_l.adiciona(instancias_selecionadas)
20 |  end
21 end
```

Após a seleção do especialista, é iniciado o processo de classificação do *self-training* (linhas 9-12, 20). A proposta deste artigo reside nas linhas 13-18, no momento que não foi possível selecionar novas instâncias para a inclusão. Quando isso acontece, é calculado a métrica silhouette para cada instância em relação à classe que ela está rotulada, considerando apenas as instâncias que o classificador especialista rotulou (linha 15). Subsequentemente, a reavaliação das instância é feita comparando o valor do silhouette

com o limiar de reavaliação (linha 16). Após isso, o comitê, treinado apenas com as instâncias rotuladas, reclassificará as instâncias que estão aptas a reavaliação (linha 17). Por fim, na linha 18, os rótulos destas instâncias são alterados no conjunto rotulado (D_l).

4.1. Configuração Experimental

Para avaliar a eficiência e viabilidade do algoritmo proposto, foi adotada uma abordagem de análise empírica. A seguir, será explicado em detalhes a configuração experimental, incluindo as características das bases de dados utilizadas e o *design* dos experimentos. Neste estudo foram utilizadas 18 bases de dados, disponíveis em repositórios e plataformas, tais como GitHub, UCI Machine Learning e Kaggle datasets. Elas foram escolhidas por sua diversidade em domínios, classes e balanceamento, permitindo avaliar a robustez e generalização da proposta. Suas características são resumidas na Tabela 1: na coluna um são descritos os nomes, a dois (#Inst) possui o número de instâncias, a três (#Att) mostra o número de atributos, a quatro (#CL) indica o número de classes, a quinta (TP) categoriza os tipos de dados ('I' = inteiro e 'R' = real) e a última (BL) classifica as bases de dados quanto ao balanceamento, indicando se são balanceadas ('B') ou desbalanceadas ('U').

Tabela 1. Propriedades das bases de dados utilizadas no experimento

Bases de Dados	#Inst	#Att	#CL	TP	BL
Blood Transfusion Service Center	748	5	2	R	U
Car	1728	22	4	I	U
Cnae-9	1080	857	9	I	B
Dermatology	366	131	6	I, R	U
Haberman	306	15	2	I, R	U
Image Segmentation	2310	19	7	I, R	B
Image Segmentation Normalized	2310	19	7	I, R	B
Iris	150	4	3	I, R	B
King-Rook vs King-Pawn	3196	40	2	I	B
Mammographic Mass	961	6	2	R	B
Multiple Features Karhunen	2000	65	10	R	B
Mushroom	8124	113	2	I	B
Ozone Level Detection	2536	73	2	R	U
Semeion	1593	257	10	I	B
Spect Heart	349	45	2	R	U
Twonorm	7400	21	2	R	B
Waveform	5000	41	3	R	B
Wilt	4839	6	2	R	U

Dado que todas as bases são originalmente rotuladas, foi possível realizar 5 configurações diferentes usando 5%, 10%, 15%, 20% and 25% de instâncias inicialmente rotuladas, conforme [Vale et al. 2021]. A escolha do conjunto de dados rotulados é realizada aleatoriamente, mas de forma estratificada e sempre garantindo um cenário típico de aprendizado semissupervisionado ($D_l \leq D_u$). Para cada base de dados foram realizados 5 vezes o *10-fold cross validation*, para garantir robustez estatística dos resultados.

Neste experimento, sete algoritmos de classificação foram utilizados para compor o ambiente experimental, garantindo diversidade e complementaridade entre os modelos,

a saber: K-Nearest Neighbors (KNN), Naive Bayes (NB), Decision Tree (DT), Random Forest (RF), XGBoost (XGB), Regressão Logística (LR) e Rede Neural Artificial (ANN). O limiar de confiança foi fixado em 95% ($threshold \geq 0,95$), como em [Vale et al. 2021].

Em relação a reavaliação de rótulos, as instâncias com valor de silhueta inferior a –0,2 foram automaticamente removidas do conjunto de dados rotulados, por serem consideradas mal posicionadas nos clusters formados. Esse limiar de -0,2 foi determinado empiricamente, após observação de seu impacto positivo na qualidade do processo de rotulagem, corroborando achados anteriores [Xie et al. 2021, Liu et al. 2024].

5. Resultados e Discussão

5.1. Análise de Performance

A Tabela 2 mostra os resultados médios das métricas analisadas, obtidos após cinco execuções independentes dos 10-fold cross validation, para cada configuração experimental. A primeira coluna, da referida tabela, indica a métrica de avaliação: acurácia (ACC), F1-score e o percentual de instâncias rotuladas ao final do processo (Instâncias Rotuladas), a segunda coluna informa o algoritmo analisado (original ou comitê com reavaliação). As colunas seguintes apresentam os resultados de acordo com o percentual de instâncias inicialmente rotuladas e os valores em negrito apontam o melhor de cada métrica comparando original e comitê com reavaliação.

Tabela 2. Resultados médios de acurácia, F1-score e instâncias rotuladas (%) para diferentes percentuais de dados inicialmente rotulados.

Métrica	Algoritmo	5%	10%	15%	20%	25%
ACC (%)	original	83.95 ± 9.57	86.06 ± 8.35	89.09 ± 8.59	**89.79 ± 8.66**	**90.92 ± 7.30**
	comitê com reavaliação	**85.93 ± 8.88**	**88.65 ± 8.12**	**89.68 ± 7.15**	89.73 ± 8.61	90.36 ± 8.81
F1-score (%)	original	72.31 ± 18.50	79.00 ± 17.47	80.07 ± 18.13	**82.36 ± 16.64**	**83.69 ± 15.81**
	comitê com reavaliação	**73.71 ± 17.78**	**80.36 ± 15.81**	**81.62 ± 15.80**	82.02 ± 17.39	83.50 ± 16.84
Instâncias Rotuladas (%)	original	61.93	67.79	68.98	73.83	76.41
	comitê com reavaliação	**62.24**	**68.49**	**69.17**	**74.16**	**77.01**

Conforme a Tabela 2, em todos os percentuais avaliados, o comitê com reavaliação de rótulos superou o *self-training* original em termos de acurácia e F1-score em 60% dos casos (3 de 5) e rotulou mais instâncias em todos os casos. A análise demonstra que o método proposto gerou um ganho médio consistente de cerca de 2 pontos percentuais na acurácia nos menores percentuais de rotulagem (5% e 10%), cenários tradicionalmente mais desafiadores para métodos semissupervisionados. Por exemplo, com 5% dos dados inicialmente rotulados, a acurácia média subiu de 83.95% para 85.93% e a F1-score de 72.31% para 73.71%. Esse padrão se repete para as demais proporções, embora com ganhos absolutos mais discretos à medida que a quantidade de dados rotulados cresce.

O impacto positivo da reavaliação de pseudo-rótulos foi evidente nos cenários com menor quantidade de dados rotulados inicialmente, onde o risco de propagação de erros é mais acentuado. Por outro lado, nas configurações com maior percentual (20% e 25%), os resultados foram mais equilibrados. Por exemplo, observando a acurácia usando 20% dos dados inicialmente rotulados o comitê com reavaliação apresentou desempenho

praticamente equivalente ao original, com acurácia de 89.73% e 89.79%, respectivamente. Este comportamento pode ser explicado pelo fato de que, com mais dados rotulados inicialmente, a necessidade de correção via reavaliação diminui, uma vez que o classificador especialista tende a ser mais confiável desde o início. Este resultado reforça a eficácia da métrica de silhueta para identificar instâncias mal agrupadas e potencialmente incorretas.

O percentual de instâncias rotuladas ao final do processo também foi ligeiramente superior no comitê com reavaliação em todas as configurações, demonstrando que a combinação da métrica de silhueta com o comitê de classificadores não apenas aprimorou a qualidade das predições, mas também potencializou a expansão segura do conjunto rotulado. Além disso, verificou-se uma redução do desvio padrão na maioria dos casos, indicando maior estabilidade do método proposto ao longo das várias execuções.

Aprofundando a comparação entre o desempenho dos métodos, foram gerados gráficos do tipo boxplot das métricas de acurácia e F1-score para as diferentes proporções de dados inicialmente rotulados (5%, 10%, 15%, 20%, 25%).

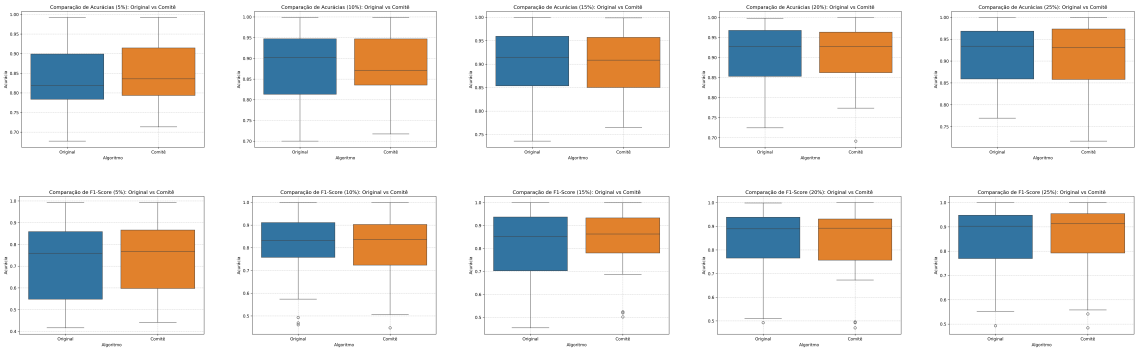

Figura 1. Gráficos boxplot de acc e F1-score

A Figura 1 mostra que, nas proporções mais baixas (5%, 10% e 15%), o comitê com reavaliação apresentou mediana superior e menor dispersão em comparação ao original, tanto em acurácia quanto em F1-score. Isso indica não apenas maior desempenho médio, mas também maior consistência entre execuções. Ademais, em vários casos, o método proposto apresentou menos *outliers*, o que sugere uma redução na ocorrência de desempenhos ruins, característicos de erros propagados por pseudo-rótulos incorretos.

À medida que o percentual de dados rotulados sobe (20% e 25%), a diferença entre os métodos diminui, com desempenhos mais próximos. Isso é coerente com a hipótese de que, com mais dados rotulados inicialmente, o classificador especialista tende a ser mais confiável e a necessidade de intervenção da métrica de silhueta e do comitê se reduz. Em geral, a análise estatística visual, reforçada pelas medidas de tendência central e dispersão, sugere que a proposta oferece ganhos expressivos em cenários com menor supervisão inicial, justamente os mais desafiadores no contexto de SSL no mundo real.

6. Conclusão e Trabalhos Futuros

Neste trabalho, foi proposta uma extensão do *self-training* com um mecanismo de reavaliação de pseudo-rótulos baseado na métrica de silhueta, aliado ao uso de um comitê de classificadores com votação ponderada. O objetivo geral foi mitigar a propagação de erros oriundos de pseudo-rótulos incorretos, um problema recorrente em cenários de SSL

com poucos dados rotulados. O algoritmo desenvolvido atua de forma iterativa, reavaliando periodicamente a qualidade dos pseudo-rótulos atribuídos e removendo aqueles associados a instâncias mal agrupadas. Quando o classificador especialista não encontra instâncias com confiança suficiente, o comitê é acionado para atribuição de novos rótulos às instâncias já pseudo-rotuladas pelo especialista, aumentando a robustez da decisão.

Os experimentos realizados em 18 bases de dados de diferentes domínios demonstraram que a proposta apresenta desempenho superior ao *self-training* original, especialmente em contextos com menor percentual de dados inicialmente rotulados. Tanto a acurácia quanto o F1-score apresentaram melhorias consistentes, além de uma leve ampliação na quantidade de instâncias rotuladas ao longo do processo. Também foi observada maior estabilidade nos resultados, evidenciada pela redução do desvio padrão em diversas configurações. A análise por meio da comparação visual com boxplots, reforçou essas observações ao mostrar melhor mediana e menor variabilidade em vários cenários. No entanto, os ganhos tendem a se tornar mais discretos à medida que aumenta a quantidade de dados rotulados inicialmente, o que é esperado, dado que o modelo original passa a ter mais confiança e menor dependência de mecanismos de correção.

O estudo revela a importância de incorporar mecanismos dinâmicos de verificação e correção no processo de pseudo-rotulagem, contribuindo para a construção de modelos semissupervisionados mais robustos e menos suscetíveis a erros acumulativos. Como trabalhos futuros, é possível investigar estratégias de ajuste adaptativo do threshold de confiança, explorar outras métricas de qualidade de cluster e avaliar a aplicabilidade da abordagem em contextos com alto desbalanceamento de classes ou em modelos baseados em aprendizado profundo. Além disso, recomenda-se a realização de testes estatísticos que permitam validar a proposta de maneira mais rigorosa e a análise do custo computacional (código-fonte: `https://github.com/labican-ufrn/semi-supervised-learning`).

Referências

Amini, M.-R., Feofanov, V., Pauletto, L., Hadjadj, L., Émilie Devijver, and Maximov, Y. (2025). Self-training: A survey. *Neurocomputing*, 616:128904.

Chapelle, O., Scholkopf, B., and Zien, A. (2006). *Semi-Supervised Learning*. The MIT Press.

Dinh, D.-T., Fujinami, T., and Huynh, V.-N. (2025). Estimating the optimal number of clusters in categorical data clustering by silhouette coefficient. *arXiv preprint arXiv:2501.15542*.

Fang, B., Li, X., Han, G., and He, J. (2023). Rethinking pseudo-labeling for semi-supervised facial expression recognition with contrastive self-supervised learning. *IEEE Access*, 11:45547–45558.

Gomes, H. M., Grzenda, M., Mello, R., Read, J., Nguyen, M.-H. L., and Bifet, A. (2021). A survey on semi-supervised learning for delayed partially labelled data streams. *ACM Computing Surveys*, 55:1 – 42.

Guérin, A., Chauvet, P., and Saubion, F. (2024). A survey on recent advances in self-organizing maps. *arXiv preprint arXiv:2501.08416*.

He, Y., Chen, W., Liang, K., Tan, Y., Liang, Z., and Guo, Y. (2023). Pseudo-label correction and learning for semi-supervised object detection. *arXiv preprint arXiv:2303.02998*.

Li, D., Liu, Z., Armaghani, D. J., Xiao, P., and Zhou, J. (2022). Novel ensemble intelligence methodologies for rockburst assessment in complex and variable environments. *Scientific Reports*, 12.

Li, J., Xie, Q., Dai, Z., Hovy, E., Le, Q. V., and Luong, M.-T. (2021). Confidence-aware pseudo label selection for semi-supervised learning. In *International Conference on Learning Representations (ICLR)*. arXiv preprint arXiv:2006.10807.

Li, X., Grandvalet, Y., and Davoine, F. (2019). Learning to self-train for semi-supervised few-shot classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 10276–10286.

Liu, Y., Zhan, L., Feng, Y., Si, P., Jiang, S., Zhao, Q., and Yan, C. (2024). Loose-tight cluster regularization for unsupervised person re-identification. *The Visual Computer*, pages 1–14. Early online version.

Oymak, S. and Gulcu, T. C. (2020). Statistical and algorithmic insights for semi-supervised learning with self-training. *arXiv preprint arXiv:2006.11006*.

Radosavovic, I., Dollar, P., Girshick, R., Gkioxari, G., and He, K. (2021). Designing pseudo-labeling for semi-supervised learning. In *International Conference on Learning Representations (ICLR)*.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Slack, D., Hilgard, S., Wu, X., Singh, S., and Talwalkar, A. (2020). Noisy student training for robust semi-supervised learning. arXiv preprint arXiv:2006.06855.

Sun, C., Shrivastava, A., Singh, S., and Gupta, A. (2021). Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 843–852.

Vale, K. M. O., Gorgônio, A. C., Flavius Da Luz, E. G., and Canuto, A. M. D. P. (2021). An efficient approach to select instances in self-training and co-training semi-supervised methods. *IEEE Access*, 10:7254–7276.

Wang, K., Zhang, C., Geng, Y., and Hu, H. (2023). Evidential pseudo-label ensemble for semi-supervised classification. *Pattern Recognition Letters*, 177:135–141.

Xie, L., Singh, A., and Precup, D. (2021). Revisiting k-means: New algorithms via bayesian nonparametrics. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 11399–11408. PMLR.

Xie, Q., Luong, M.-T., Hovy, E., and Le, Q. V. (2020). Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10687–10698.

Zhu, X. and Goldberg, A. B. (2009). *Introduction to Semi-Supervised Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool.