

Enriquecimento de Dados com Base em Estatísticas de Grafo de Similaridade para Melhorar o Desempenho em Modelos de ML Supervisionados de Classificação

Ney Barchilón¹, Hélio Côrtes Vieira Lopes¹, Marcos Kalinowski¹, Jefry Sastre Perez¹

¹Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), Rio de Janeiro, RJ, Brasil
Departamento de Informática

nbarchilon@inf.puc-rio.br, lopes@inf.puc-rio.br,
kalinowski@inf.puc-rio.br, jefry.sastre@gmail.com

Abstract. *This research proposes a method for enriching tabular datasets using graph statistics, aiming to improve the performance of supervised ML classification models. The method builds a graph based on the similarity between instances in the dataset and extracts features from the graph to enrich the original dataset. Evaluated on 10 public datasets from different areas of knowledge, with 7 machine learning models, the method provided an average increase of 4.9% in accuracy. The results demonstrate the effectiveness of the method as an alternative to improve model performance in scenarios where data sets lack the characteristics necessary for traditional enrichment approaches using graphs.*

Resumo. *Esta pesquisa propõe um método para o enriquecimento de conjuntos de dados tabulares utilizando estatísticas de grafo, visando melhorar o desempenho de modelos de ML supervisionados de classificação. O método constrói um grafo a partir da similaridade entre as instâncias do conjunto de dados e extrai características do grafo para enriquecer o conjunto de dados original. Avaliado em 10 conjuntos de dados públicos de diferentes áreas do conhecimento, com 7 modelos de aprendizado de máquina, o método proporcionou um aumento médio de 4,9% na acurácia. Os resultados demonstram a efetividade do método como uma alternativa para melhorar o desempenho de modelos em cenários que conjuntos de dados carecem das características necessárias para as abordagens tradicionais de enriquecimento com a utilização de grafo.*

1. Introdução

A otimização do desempenho dos modelos de aprendizado de máquina supervisionados representa um desafio constante, especialmente em contextos com conjuntos de dados de alta dimensionalidade ou com numerosos atributos correlacionados, mas depende da qualidade dos dados. O enriquecimento de dados aprimora essa qualidade, fornecendo informações adicionais para modelos de aprendizado de máquina [Dong and Oyamada 2022]. Trabalhos anteriores [Abdelmageed 2020] e [Dong and Oyamada 2022] demonstraram que o enriquecimento de dados por meio de grafos pode melhorar a performance dos modelos de predição, integrando informações fundamentais sobre os dados.

A presente pesquisa propõe um método para enriquecer conjuntos de dados tabulares visando otimizar modelos de classificação. O método baseia-se na extração de

estatísticas de um grafo construído a partir da similaridade entre as instâncias do conjunto de dados. Trabalhos como os de [Gulum 2018] e [Alharbi and Alsubhi 2021] empregam conjuntos de dados tabulares, que já contêm informações sobre entidades e relacionamentos, para construir seus grafos de conhecimento. Eles então extraem estatísticas desses grafos para enriquecer o conjunto de dados original, preparando-o para ser utilizado em modelos de aprendizado de máquina.

No entanto, essas abordagens convencionais dependem da identificação de entidades e relacionamentos no conjunto de dados original, o que pode ser limitante em cenários onde tais características não existem. Os estudos de [Zaki et al. 2021] e [Albreiki et al. 2023] adotam uma abordagem que envolve a utilização de estatísticas derivadas do grafo para enriquecer conjuntos de dados tabulares, preparando-os para a aplicação em modelos de aprendizado de máquina. Esses pesquisadores recorrem a métodos alternativos para estabelecer conexões (arestas) na construção do grafo, como a captura de correlações estruturais entre os dados e a definição de distâncias entre as instâncias.

O objetivo geral deste trabalho é propor e avaliar uma solução flexível e eficaz para o enriquecimento de conjuntos de dados tabulares baseado em estatísticas extraídas de um grafo construído a partir da similaridade entre as instâncias do conjunto de dados, considerando distintamente as características categóricas e numéricas, visando melhorar a performance de modelos de classificação. As questões de pesquisa abordam como trabalhar conjuntos de dados para enriquecimento com estatísticas do grafo de similaridade, a eficácia do método em diferentes domínios de conhecimento com o impacto na performance dos modelos e a comparação com outras técnicas de melhoria de desempenho.

Ao remover as restrições impostas pela necessidade de identificação, entre as características do conjunto de dados, de entidades e relacionamentos para a construção do grafo e a utilização de características que já existem nesse conjunto de dados para tanto, a pesquisa visa proporcionar um método alternativo acessível, aplicável em uma variedade de cenários, e capaz de contribuir para o avanço no campo do enriquecimento de dados tabulares para ML. Além disso, permite sua integração com outras soluções de enriquecimento existentes.

Este artigo está organizado conforme segue. Na seção 2 efetuamos uma revisão da literatura, apresentando os trabalhos relacionados para situar a presente pesquisa no contexto mais amplo do enriquecimento de dados tabulares. Na seção 3, descrevemos o método proposto e as etapas do processo. A seguir, a Seção 4 apresenta os experimentos realizados e os resultados obtidos. Finalmente, a seção 5 conclui este trabalho.

2. Trabalhos Relacionados

Na área de enriquecimento de conjuntos de dados tabulares, estudos recentes têm explorado métodos para melhorar a predição de modelos de *machine learning* (ML) por meio da incorporação de dados externos. [Dong and Oyamada 2022] propuseram um método automatizado de enriquecimento de dados tabulares com colunas externas provenientes de data lakes, resultando em melhorias significativas na predição de modelos de ML. Outros estudos investigaram a pesquisa por tabelas em data lakes, como o framework PEXESO proposto por [Dong et al. 2020], que utiliza abordagens baseadas em similaridade de alta dimensão para descobrir tabelas que podem ser unidas de forma efici-

ente e significativa, e abordagens baseadas em grafos para melhorar a correspondência e interpretação semântica de tabelas, conforme explorado por [Jiomekong and Foko 2022] e [Gottschalk and Demidova 2022].

Estudos recentes também abordam a extração de estatísticas de grafos de conhecimento a partir de dados tabulares para enriquecer conjuntos de dados originais. Por exemplo, [Sanz and Duarte 2019] propuseram uma abordagem para detecção de intrusão em redes virtuais, enquanto Baumann et al. (2017) [Baumann et al. 2017] exploram a construção de grafos de sessões de usuário na web com base em dados de clickstream. Gulum [Gulum 2018] investiga o uso de características extraídas de grafos para prever resultados de corridas de cavalos. Na mesma linha, [Alharbi and Alsubhi 2021] propuseram um modelo de detecção de botnets baseado em aprendizado de máquina que incorpora características extraídas de grafos construídos a partir de fluxos de rede.

Esses estudos demonstram a aplicabilidade e eficácia da extração de estatísticas de grafos de conhecimento para enriquecer conjuntos de dados em diversas aplicações, desde segurança de rede até previsão de resultados e detecção de padrões. No entanto, essas abordagens convencionais dependem da identificação de entidades e relacionamentos no conjunto de dados original, o que pode ser limitante em cenários onde tais características não são explícitas.

Nessa perspectiva, os trabalhos de [Zaki et al. 2021] e [Albreiki et al. 2023] estão alinhados com o método proposto nesse estudo. Em um estudo no campo da saúde [Zaki et al. 2021] propõem uma abordagem alternativa, baseada em correlação para converter dados tabulados em grafos de conhecimento, capturando correlações estruturais entre os dados e melhorando o desempenho de modelos de classificação em conjuntos de dados de saúde. Por outro lado, em um contexto educacional, [Albreiki et al. 2023] investigam a identificação de estudantes em risco de desempenho acadêmico usando características topológicas extraídas de grafos construídos a partir das medições das distâncias das instâncias dos dados para estabelecer conexões (arestas) na construção do grafo, preparando-os para a aplicação em modelos de aprendizado de máquina (ML). Os resultados de ambos os trabalhos demonstram a eficácia dessa abordagem em conjuntos de dados tabulares, sem a identificação de características específicas do conjunto de dados para determinação de vértices e estabelecimento das arestas do grafo.

Como será visto adiante, em contraste com os trabalhos de [Zaki et al. 2021] e [Albreiki et al. 2023], que aplicaram a validação cruzada após a construção do grafo com todo o conjunto de dados, o presente estudo utilizou o método *holdout* para a separação dos conjuntos de treino e teste, construindo os grafos de treino e teste separadamente. Outra diferença significativa é o tratamento distinto das similaridades de características categóricas e numéricas.

3. Metodologia

O *pipeline* do método proposto está alinhado ao *pipeline* padrão para a predição de modelos de ML supervisionados como aquele apresentado por [Escovedo and Koshiyama 2020] para projetos de ciência de dados, mas com algumas particularidades. A Figura 1 representa uma visão geral do método proposto. São seis grandes etapas: Tratamento de Dados (1), Treino e Teste Modelos FO (2), Construção Grafo de Treino (3), Construção Grafo de Teste (4), Treino e Teste Modelos

Grafo (5) e Avaliação Resultados com *Benchmark* (6).

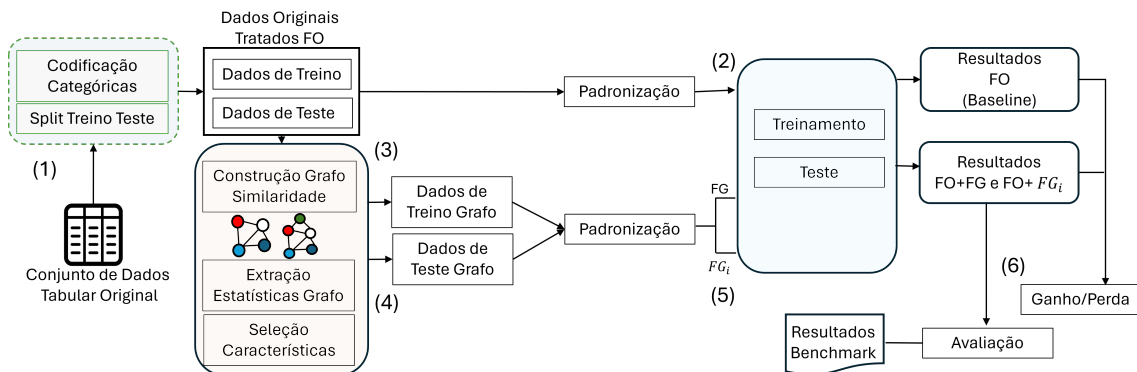


Figura 1. Visão geral da metodologia sugerida com suas etapas

A etapa de Tratamento de Dados (1) visa aprimorar a qualidade dos dados para os modelos antes de seu enriquecimento. As características categóricas são codificadas e o conjunto de dados é então dividido em treino e teste, garantindo que os dados de treino não tenham contato com os dados de teste. As características numéricas são escalonadas utilizando a padronização. No conjunto de dados PIMA as colunas com valores ausentes receberam a média dos valores preenchidos e nos demais conjuntos as colunas com valores ausentes foram desconsideradas. Nenhum outro procedimento de tratamento foi aplicado ao conjunto de dados, a fim de evitar influências no processo de comparação de resultados.

Na etapa (2), os modelos são treinados com os 'Dados de Treino' e 'Dados de Teste', a fim de estabelecer uma linha de base (*baseline*) para comparação com resultados obtidos após enriquecimento dos dados. Para evitar vieses nos resultados, os algoritmos são treinados com parâmetros padrão. Essa abordagem garante que a única variável entre os testes seja o enriquecimento dos dados com as estatísticas extraídas do grafo.

A etapa (3) é onde são efetuados os procedimentos para construção do grafo de treino, tendo como base o conjunto de 'Dados de Treino'. Cada instância é representada por um vértice no grafo e, conforme a Figura 2, as características categóricas e numéricas são separadas em subconjuntos distintos. A similaridade entre os objetos em cada subconjunto é medida utilizando técnicas adequadas para cada tipo de característica. Pares de vértices com similaridade acima de um limite (threshold) são conectados por arestas. Esse limite foi definido como a mediana dos valores de proximidade entre um vértice e os demais. O peso da aresta é a soma das similaridades entre as características categóricas e numéricas. Onze estatísticas derivadas do grafo de treino são extraídas para cada vértice. Dessas são selecionadas seis estatísticas pela técnica de filtro [Saeys et al. 2007]. As estatísticas selecionadas são incorporadas ao conjunto de 'Dados de Treino' como novas características, criando o conjunto 'Dados de Treino Grafo'.

A Figura 3 apresenta um exemplo hipotético simples da construção do grafo de Treino utilizando um conjunto de dados com 3 instâncias e 6 colunas (3 categ. e 3 num.) e medidas de proximidade *Overlap* (categ.) e *Manhattan* (num.).

O procedimento para construção do grafo de teste é efetuado na etapa (4), tendo como base 'Dados de Teste' e o grafo de treino. O grafo de teste é inicialmente igual ao

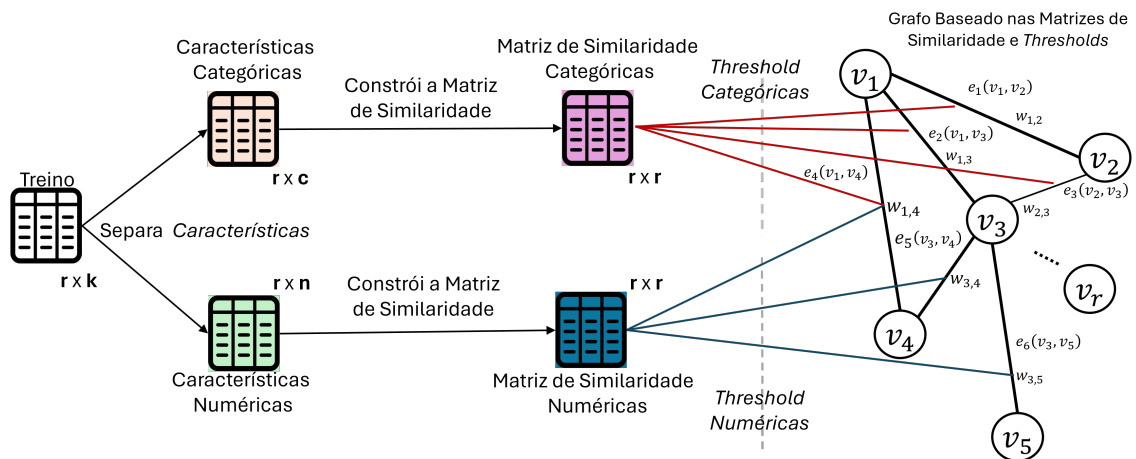


Figura 2. Construção do Grafo de Treino pela Similaridade

grafo de treino. Os dados do conjunto de teste são incluídos no grafo de teste utilizando os mesmos critérios de similaridade do grafo de treino. As mesmas estatísticas de grafo extraídas e selecionadas do grafo de treino são extraídas do grafo de teste e incorporadas ao conjunto de dados de teste tratado, criando o conjunto 'Dados de Teste Grafo'.

Na etapa (5), os modelos de predição selecionados são treinados utilizando os conjuntos de dados enriquecidos com as estatísticas extraídas dos grafos de treino e teste. Antes do treinamento, as características numéricas extraídas dos grafos são Padronizadas. Os modelos e parâmetros utilizados são os mesmos da etapa (2). Dois tipos de submissão de dados serão utilizados para treinar e testar os modelos de aprendizado de máquina: Primeiro o conjunto completo (FO+FG) que inclui todas as características originais (FO) e as seis estatísticas selecionadas do grafo de similaridade (FG). O segundo com subconjuntos (FO+FG_i) que incluem as características originais (FO) e uma das novas características extraídas e selecionadas do grafo de similaridade (FG_i é uma estatística do grafo selecionada, onde $i=1, \dots, 6$), resultando em subconjuntos individuais com cada uma das novas características extraídas e selecionadas do grafo de similaridade.

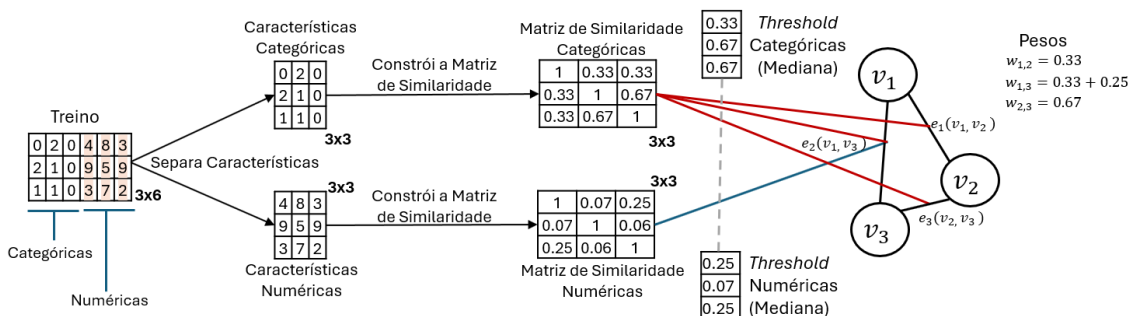


Figura 3. Exemplo da Construção do Grafo de Treino pela Similaridade

Os resultados dos testes das etapas (2) e (5) são comparados na etapa (6). O objetivo é avaliar se a incorporação das estatísticas extraídas do grafo nos conjuntos de dados de teste contribuiu para a melhora do desempenho dos modelos de predição.

4. Experimento

Foram selecionados dez conjuntos de dados públicos abrangendo uma diversidade de domínios de conhecimento, variando em dimensões, tamanhos e perfis de balanceamento de classe, conforme Tabela 1. O conjunto de dados Pima Indians *Dataset* foi obtido no repositório da [Kaggle](https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database)¹ e os demais no repositório da [UCI Machine Learning Repository](https://www.archive.ics.uci.edu)².

Tabela 1. Resumo dos Conjuntos de Dados

Sigla	Dataset	Domínio	Linhas	Cols	Cat.	Num.	Label	Bal.	Teste
BCC	Breast Cancer Coimbra	Saúde	116	9	0	9	B	S	20%
CAR	CAR Evaluation	Auto	1.728	6	6	0	M	N	30%
DM	Dermatology	Saúde	366	34	33	1	M	N	23%
HDH	Heart Disease Hungarian	Saúde	294	13	9	4	B	S	30%
HDHNI	Heart Disease Hungarian Non-Invasive	Saúde	294	2	2	0	B	S	30%
PIMA	Pima Indians Diabetes	Saúde	768	8	1	7	B	N	30%
VCB	Vertebral Column	Saúde	310	6	0	6	B	N	50%
VCM	Vertebral Column Binary	Saúde	310	6	0	6	M	N	25%
WM	Wine	Bebida	178	13	0	13	M	S	20%
WQ	Wine Quality	Bebida	1.599	11	0	11	M	N	30%

As medidas de proximidade utilizadas foram Jaccard [Han et al. 2012], Overlap [Sulc and Řezankova 2014] e Eskin [Sulc and Řezankova 2014] para características categóricas e para numéricas Euclidean [Gan et al. 2007], Cosine [Tan et al. 2019], Manhattan [Gan et al. 2007] e Mahalanobis [Gan et al. 2007].

As onze estatísticas de nível nodal (vértices) selecionadas para análise foram: Medidas de Centralidade: Degree(DG) [Newman 2018], Degree Centrality(DC) [Srinivasan et al. 2020], Average Neighbor Degree(AN) [Barrat et al. 2004], Betweenness Centrality(BC) [Brandes 2001], Closeness Centrality(CC) [Freeman 1977], Eigenvector Centrality(EC) [Newman 2018], Load Centrality(LC) [Newman 2005]; Medidas de Estrutura Local: Clustering(CL) [Onnela et al. 2005] e Triangles(TR) [Needham and Hodler 2019]; Medidas de Análise de Links: Page-rank(PR) [Brin and Page 1998] e Hits(HT) [Langville and Meyer 2004].

Os algoritmos de classificação trabalhados foram: Naïve Bayes (NBG), Máquina de Vetores de Suporte (SVM), Regressão Logística (RLOG), Árvore de Decisão (DT) e Label Propagation (LPA). Pelo lado dos algoritmos Ensemble de Bagging e Boosting temos Florestas Aleatórias (RFOR) e Extreme Gradient Boosting (XGB), respectivamente.

Para avaliar o desempenho da metodologia proposta, empregamos a separação de treino e teste *holdout* [Kuncheva 2004], isto é, o conjunto de dados original será separado entre um conjunto de treino e teste, com uma proporção orientada pela mesma proporção utilizada em um artigo de referência daquele conjunto de dados. Medidas de avaliação amplamente aceitas foram empregadas para avaliação das classificações, incluindo Acurácia, Precisão, Sensibilidade (Recall) e F1.

Dois cenários foram utilizados para avaliar o desempenho dos modelos de predição: O Cenário 1 com parâmetros padrão nos modelos; Cenário 2 com uma grade

¹www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

²www.archive.ics.uci.edu

de parâmetros (Tabela 2) nos modelos. Os cenários seguem o paradigma "situação anterior/posterior" para gerar resultados para avaliação, permitindo investigar as mudanças nas medidas de desempenho dos modelos antes/depois da adição de características do grafo.

Para cada conjunto de dados, em ambos os cenários, o pipeline do método foi processado dez vezes e a média destes foi adotada como resultado.

Tabela 2. Hyperparâmetros por Modelo

Modelo	Hiperparâmetros para Grid
DT	criterion['gini', 'entropy'], max_depth[5, 10], min_samples_leaf[1, 2, 3], splitter['best', 'random'], max_features['sqrt', 'log2']
LPA	kernel["knn", "rbf"], n_neighbors [3, 5, 7, 9, 11]
NBG	Sempre utilizado com padrão
RFOR	n_estimators[50, 100], criterion['gini', 'entropy'], max_depth[5, None], min_samples_leaf [1, 5], max_features['sqrt', 'log2']
RLOG	solver['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'], c_values[0.01, 0.1, 1]
SVM	c_values[0.001, 0.1, 1.0, 10], kernel_values['linear', 'poly', 'rbf'], gamma=['scale', 'auto']
XGB	max_depth[3, 6, 10], n_estimators[50, 100], learning_rate[0.01, 0.1, 0.3, 0.4]

Resultados

A Tabela 3 apresenta os ganhos/perdas de Acurácia com a inclusão no conjunto de dados original das estatísticas extraídas do grafo. Por exemplo, no conjunto de dados BCC, para o Cenário 1 (C1: com parametros padrão no modelo), no modelo SVM, o melhor resultado com ganho de Acurácia foi de 16,7%. A média de melhor resultado de Acurácia no conjunto de dados BCC, para o Cenário 1 (C1) foi de 14,3%. A média de ganhos, em ambos os cenários, foi de 4,9%. Os resultados agregados nos dois Cenários indicam que houve apenas diferenças pontuais entre os dois cenários. Observa-se que a maioria dos conjuntos de dados e modelos experimentou melhorias nas medidas de desempenho após a inclusão das características do grafo, com destaque para os modelos DT, NBG, RFOR, SVM e XGB, que alcançaram, em média, os maiores ganhos.

A presença de um maior número de características categóricas limitou os ganhos na métrica Acurácia com a inclusão das estatísticas de grafo, como observado nos conjuntos de dados HDHNI e CAR. Isso sugere que algumas medidas de similaridade adotadas para características categóricas (*Jaccard*, *Overlap* e *Eskin*) podem não ser adequadas para estabelecer as arestas do grafo, tornando as características extraídas do grafo menos informativas para o modelo. Medidas de similaridade que tratam todas as diferenças igualmente, independentemente da ordem, não conseguem diferenciar adequadamente valores próximos ou distantes na escala ordinal, resultando em uma representação pouco adequada das relações entre as instâncias. Em contraste, conjuntos de dados com muitas características numéricas, como BCC, VCB e VCM, mostram maiores ganhos, sugerindo que as similaridades e estatísticas do grafo selecionadas são mais eficazes nesse contexto.

Os gráficos na Figura 4, posicionados lado a lado, ilustram a frequência (eixo x) e o desempenho (%) na métrica Acurácia (eixo y) das medidas de similaridade (gráfico 1) e das estatísticas do grafo (gráfico 2) nos experimentos realizados com a inclusão no conjunto de dados original das estatísticas extraídas do grafo.

Pelo lado das similaridades (Figura 4 - gráfico 1), verificamos o destaque da medida *Jaccard* que teve a maior frequência entre os melhores resultados para métrica Acurácia entre as medidas de similaridade para características categóricas, embora com uma pequena média de ganho. Uma vez que a medida *Jaccard* trabalha melhor ca-

Tabela 3. Cenário 1 e 2: Ganho/Perda Acurácia (%) por Dataset e Modelo

Dataset	Cenário*	DT	LPA	NBG	RFOR	RLOG	SVM	XGB	Média
BCC	C1	20,8	8,4	25,0	12,5	4,1	16,7	12,5	14,3
	C2	29,2	8,4	25,0	12,5	8,3	20,8	16,6	17,3
CAR	C1	0,0	0,8	9,1	-1,7	0,0	-0,8	-0,5	1,0
	C2	0,9	0,9	7,8	-1,5	0,0	-0,3	-2,2	0,8
DM	C1	0,0	2,3	1,1	2,3	2,3	0,0	1,2	1,3
	C2	1,2	2,3	0,0	2,3	1,1	1,2	1,1	1,3
HDH	C1	0,0	1,2	1,2	3,4	1,1	1,1	2,3	1,5
	C2	2,3	1,2	1,2	5,6	1,1	1,2	0,0	1,8
HDHNI	C1	0,0	1,1	0,0	2,2	0,0	1,1	0,0	0,6
	C2	1,1	1,1	1,1	1,1	0,0	0,0	4,5	1,3
PIMA	C1	4,3	6,5	2,6	3,5	2,6	2,1	1,3	3,3
	C2	3,9	1,3	2,6	2,2	2,6	2,2	3,0	2,5
VCB	C1	16,2	9,6	10,3	14,8	11,0	10,9	13,5	12,3
	C2	16,2	5,1	0,0	13,5	10,3	11,6	13,5	10,0
VCM	C1	17,9	7,7	12,8	15,4	5,1	10,2	12,8	11,7
	C2	10,3	7,7	12,8	14,1	5,1	14,1	11,5	10,8
WM	C1	0,0	0,0	2,8	0,0	2,8	2,8	2,8	1,6
	C2	2,8	0,0	2,8	0,0	0,0	0,0	0,0	0,8
WQ	C1	0,6	2,1	2,1	2,9	1,7	1,0	1,9	1,8
	C2	4,2	2,1	1,9	2,8	1,7	1,8	1,7	2,3
Média	C1	6,0	4,0	6,7	5,5	3,1	4,5	4,8	4,9
	C2	7,2	3,0	5,5	5,3	3,0	5,3	5,0	4,9

* C1 e C2: Cenário 1 e 2

racterísticas binárias, esse resultado sugere que a transformação de codificação binária dos domínios de características categóricas ajudou essa medida na captação de similaridade entre as instâncias dos conjuntos dados. Para características numéricas, observa-se a concentração dos melhores resultados da métrica Acurácia nas medidas *Euclidean* e *Mahalanobis*. Tanto a *Euclidean* quanto a *Mahalanobis* são medidas de distância que avaliam a proximidade entre pontos de dados no espaço de características e que podem ser particularmente influenciadas se as características extraídas do grafo se baseiam principalmente na proximidade ou na distância entre as instâncias, então essas medidas podem capturar com eficácia tais relações.

Entre as estatísticas extraídas do grafo (Figura 4 - gráfico 2), as mais relevantes para os ganhos de performance foram *degree* (DG), *eigenvector centrality* (EC), *clustering* (CL), *hits* (HT) e o conjunto de todas as estatísticas (FG). Destaca-se o tempo de processamento das estatísticas CL, BC e LC (gráfico na Figura 5), que juntas consomem cerca de 97% do tempo de extração como reflexo da complexidade de tempo $O(n.m + n^2 \log n)$, $O((m + n).n^2)$ e $O(n^3)$, respectivamente, onde n é o número de vértices e m o número de arestas no grafo. Observa-se ainda que, no gráfico na Figura 6, apenas CL gerou resultados mais expressivos no Cenário 2, enquanto BC e LC contribuíram menos, apesar do alto consumo de processamento.

Uma forma de contornar o alto custo computacional de algumas estatísticas seria aumentar o limite (*threshold*) para o estabelecimento de arestas no grafo, tornando-o

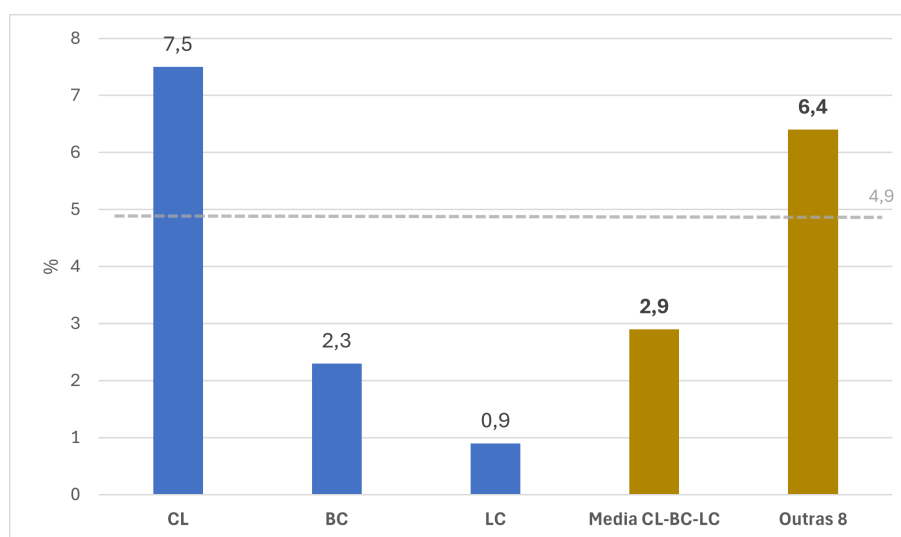


Figura 6. Ganho Médio no Cenário 2 por Estatística do Grafo – Acurácia (%)

modelos, sugerem que o método proposto tem o potencial de ser mais uma alternativa competitiva de enriquecimento de conjuntos de dados tabulares para melhorar o desempenho de modelos de ML supervisionados de classificação.

5. Conclusão

Contribuições

Este estudo oferece uma metodologia para o enriquecimento de conjuntos de dados tabulares, utilizando estatísticas de grafo de similaridade para melhorar o desempenho de modelos de aprendizado de máquina supervisionados de classificação. Os resultados experimentais sugerem que a inclusão de estatísticas de grafo pode melhorar a acurácia dos modelos, proporcionando ganhos médios de 4,9% nos cenários avaliados. A comparação com outros estudos anteriores, sobre os mesmos conjuntos de dados, revelou resultados positivos na maioria dos casos, reforçando a competitividade do método proposto. Além disso, identificamos modelos, medidas de similaridade e estatísticas do grafo que se destacaram, com *insights* para aplicações futuras nessa área.

Limitações

Observamos que conjuntos de dados com características categóricas apresentaram desempenho inferior ao incluir estatísticas de grafo de similaridade. Além disso, identificamos um aumento no tempo de processamento, devido ao cálculo das estatísticas do grafo BC, LC e CL, ressaltando a importância de considerar cuidadosamente a complexidade de tempo na seleção das estatísticas do grafo ao aplicar essa metodologia.

Trabalhos Futuros

Uma área promissora para pesquisa é a avaliação de novas medidas de similaridade e dissimilaridade, adaptadas para diferentes tipos de atributos, que pode enriquecer a compreensão das relações entre instâncias e aprimorar a qualidade do grafo gerado, das estatísticas extraídas e dos modelos de predição. Além disso, a expansão do número de estatísticas extraídas do grafo para capturar a estrutura dos relacionamentos e o refinamento dos procedimentos para calcular os pesos dos relacionamentos são fundamentais para melhorar a precisão das informações extraídas e a relevância para os modelos de ML. Da mesma forma, explorar estatísticas de nível global do grafo permitiriam uma compre-

Tabela 4. Comparação com Trabalhos de Pesquisas Anteriores

Dataset	Autor	Método/Modelo/Acurácia(%)	Método Proposto Acurácia (%)
BCC	[Naveen et al. 2019] [Alfian et al. 2022]	Treino/Teste(%): 80/20*. Bagging KNN: 100% 10-fold CV. SVM/ETREE: 80,2%	SVM: 95,8
CAR	[Jalali et al. 2017] [Uzut and Buyrukoglu 2020] [Rehman et al. 2018]	Treino/Teste(%): 70/30*. EAC: 96,0% 5-fold CV. GB: 99,4% Treino/Teste(%): 90/10. EAC: 94,8%	SVM: 98,5
DM	[Sharma and Hota 2013] [Putatunda 2020] [Rathore et al. 2022]	Treino/Teste(%): 77/23*. ANN/SVM: 99,0% 10-fold CV. Derm2Vec 96,9% Treino/Teste(%): 80/20. RFOR: 97,8%	SVM: 98,8
HDH	[Garate-Escamila et al. 2020] [Saboor et al. 2022] [Bashir et al. 2021]	Treino/Teste(%): 70/30*. RFOR: 99,0%. 10-fold CV. SVM: 96,7% 10-fold CV. Ensemble(SVM+NB+AutoMLP): 83,5%	RFOR: 98,9
HDHNI	[Garate-Escamila et al. 2020]	Treino/Teste(%): 70/30*. DT 82,9%.	DT: 96,5
PIMA	[Chang et al. 2022] [Zaki et al. 2021] [Kibria 2022]	Treino/Teste(%): 70/30*. RFOR: 79,6% 10-fold CV. GCN: 99,1% Treino/Teste(%): 80/20. Votting XGB/RFOR: 90%	RFOR: 78,4
VCB	[Ansari et al. 2013] [Raihan-Al-Masud and Mondal 2020]	Treino/Teste(%): 50/50*. ANN: 92,1% Treino/Teste(%): 78/22. ANN: 87,0%	XGB 98,7
VCM	[Reshi et al. 2021] [Ramamy et al. 2020]	Treino/Teste(%): 75/25*. 1.ETREE: 99%; 2.RLOG: 93%. Treino/Teste(%): 50/50. SVM: 85,0%	RFOR: 97,4
WM	[Ojha and Nicosia 2020] [Di et al. 2020]	Treino/Teste(%): 80/20*. MONT: 100% 10-fold CV. KNN-LDS: 98%	SVM: 100
WQ	[Kumar et al. 2020] [Cardone and Di Martino 2023] [Gupta and Chandrasekaran 2021]	Treino/Teste(%): 70/30*. SVM: 68,6% Treino/Teste(%): 80/20. ANN-F1: 76% Treino/Teste(%): 75/25. MP5: 81,8%	RFOR 69,0

*Utilizado como referência na proporção de separação de treino/teste do método proposto para esse conjunto de dados

ensão mais abrangente do comportamento do grafo como um todo, o que poderia ajudar a identificar padrões de conectividade, comunidades ou centralidades globais que poderiam antecipar se as estatísticas de nível nodal e/ou de arestas levariam a melhorias significativas no desempenho dos modelos de ML quando incorporadas aos conjuntos de dados originais. Por fim, a adaptação da metodologia proposta para aplicação em modelos de regressão e modelos não supervisionados, representam perspectivas para futuras pesquisas, permitindo sua aplicação em diferentes cenários, ampliando seu alcance em diversas aplicações de ML.

Referências

- Abdelmageed, N. (2020). Towards transforming tabular datasets into knowledge graphs. In *The Semantic Web: ESWC 2020 Satellite Events: Heraklion, Crete, Greece, May 31 – June 4, 2020*, pages 217—228, Berlin, Heidelberg. Springer-Verlag.
- Albreiki, B., Habuza, T., and Zaki, N. (2023). Extracting topological features to identify at-risk students using machine learning and graph convolutional network models. *Int. J. Educ. Technol. High. Educ.*, 20(1). DOI: <https://doi.org/10.1186/s41239-023-00389-3>.
- Alfian, G., Syafrudin, M., Fahrurrozi, I., Fitriyani, N. L., Atmaji, F. T. D., Widodo, T., Bahiyah, N., Benes, F., and Rhee, J. (2022). Predicting breast cancer from risk factors using SVM and extra-trees-based feature selection method. *Computers*, 11(9):136.

- Alharbi, A. and Alsubhi, K. (2021). Botnet detection approach using graph-based machine learning. *IEEE Access*, 9:99166–99180. DOI: 10.1109/ACCESS.2021.3094183.
- Ansari, S., Sajjad, F., ul Qayyum, Z., Naveed, N., and Shafi, I. (2013). Diagnosis of vertebral column disorders using machine learning classifiers. In *2013 International Conference on Information Science and Applications, ICISA*, pages 1–6.
- Barrat, A., Barthélemy, M., Pastor-Satorras, R., and Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences*, 101(11):3747–3752.
- Bashir, S., Almazroi, A., Ashfaq, S., Almazroi, A., and Khan, F. (2021). A knowledge-based clinical decision support system utilizing an intelligent ensemble voting scheme for improved cardiovascular disease prediction. *IEEE Access*, PP:1–1.
- Baumann, A., Haupt, J., Gebert, F., and Lessmann, S. (2017). Changing perspectives: Using graph metrics to predict purchase probabilities. *Expert Systems with Applications*, 94. DOI: <https://doi.org/10.1016/j.eswa.2017.10.046>.
- Brandes, U. (2001). A faster algorithm for betweenness centrality. *The Journal of Mathematical Sociology*, 25(2):163–177.
- Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1):107–117. Proceedings of the Seventh International World Wide Web Conference.
- Cardone, B. and Di Martino, F. (2023). A novel classification algorithm based on multi-dimensional fl fuzzy transform and pca feature extraction. *Algorithms*, 16:128.
- Chang, V., Bailey, J., Xu, Q. A., and Sun, Z. (2022). Pima indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Comput. Appl.*, 35(22):1–17.
- Di, X., Yu, P., Bu, R., and Sun, M. (2020). Mutual information maximization in graph neural networks. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE. DOI: <https://doi.org/10.1109/IJCNN48605.2020.9207076>.
- Dong, Y. and Oyamada, M. (2022). Table enrichment system for machine learning. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22*, pages 3267–3271, New York, NY, USA. Association for Computing Machinery.
- Dong, Y., Takeoka, K., Xiao, C., and Oyamada, M. (2020). Efficient joinable table discovery in data lakes: A high-dimensional similarity-based approach. *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pages 456–467.
- Escovedo, T. and Koshiyama, A. (2020). *Introducao a Data Science - Algoritmos de Machine Learning e metodos de analise*. Casa doCodigo.
- Freeman, L. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 40:35–41.
- Gan, G., Ma, C., and Wu, J. (2007). *Data Clustering: Theory, Algorithms, and Applications*. ASA-SIAM Series on Statistics and Applied Probability. Society for Industrial and Applied Mathematics.

- Garate-Escamila, A. K., Hajjam El Hassani, A., and Andres, E. (2020). Classification models for heart disease prediction using feature selection and pca. *Informatics in Medicine Unlocked*, 19:100330.
- Gottschalk, S. and Demidova, E. (2022). Tab2kg: Semantic table interpretation with lightweight semantic profiles. *Semantic Web*, 13(3):571—597.
- Gulum, M. (2018). *Horse racing prediction using graph-based features*. PhD thesis.
- Gupta, M. and Chandrasekaran, V. (2021). A study and analysis of machine learning techniques in predicting wine quality. *International Journal of Recent Technology and Engineering*, 10:314–321.
- Han, J., Kamber, M., and Pei, J. (2012). *Data Mining: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science.
- Jalali, V., Leake, D., and Forouzandehmehr, N. (2017). Learning and applying case adaptation rules for classification: An ensemble approach. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*.
- Jiomekong, A. and Foko, B. (2022). Towards an approach based on knowledge graph refinement for tabular data to knowledge graph matching. *Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab)*, CEUR-WS. org.
- Kibria, H. (2022). An ensemble approach for the prediction of diabetes mellitus using a soft voting classifier with an explainable AI. ” sensors. *Sensors*, 22.
- Kumar, S., Agrawal, K., and Mandan, N. (2020). Red wine quality prediction using machine learning techniques. In *2020 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–6.
- Kuncheva, L. (2004). *Combining Pattern Classifiers: Methods and Algorithms*.
- Langville, A. and Meyer, C. (2004). A survey of eigenvector methods of web information retrieval. *SIAM Review*, 47.
- Naveen, Sharma, R. K., and Ramachandran Nair, A. (2019). Efficient breast cancer prediction using ensemble machine learning models. In *2019 4th International Conference on Recent Trends on Electronics, Information, Communication Technology (RTEICT)*, pages 100–104.
- Needham, M. and Hodler, A. (2019). *Graph Algorithms: Practical Examples in Apache Spark and Neo4j*. O’Reilly Media.
- Newman, M. (2018). *Networks*. Oxford University Press.
- Newman, M. J. (2005). A measure of betweenness centrality based on random walks. *Social Networks*, 27(1):39–54.
- Ojha, V. and Nicosia, G. (2020). Multi-objective optimisation of multi-output neural trees. *2020 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8.
- Onnela, J.-P., Saramäki, J., Kertész, J., and Kaski, K. (2005). Intensity and coherence of motifs in weighted complex networks. *Phys. Rev. E*, 71:065103.
- Putatunda, S. (2020). A hybrid deep learning approach for diagnosis of the erythematosquamous disease. In *2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, pages 1–6.

- Raihan-Al-Masud, M. and Mondal, M. R. H. (2020). Data-driven diagnosis of spinal abnormalities using feature selection and machine learning algorithms. *PLoS One*.
- Ramasamy, M., Abdulkadhar, S., and Natarajan, J. (2020). Deep neural network for the automatic classification of vertebral column disorders.
- Rathore, A. S., Arjaria, S., Gupta, M., Chaubey, G., Mishra, A., and Rajpoot, V. (2022). Erythematous-squamous diseases prediction and interpretation using explainable ai. *IETE Journal of Research*.
- Rehman, Z., Fayyaz, H., Shah, A., Aslam, N., Hanif, M., and Abbas, S. (2018). Performance evaluation of mlpnn and nb: A comparative study on car evaluation dataset.
- Reshi, A. A., Ashraf, I., Rustam, F., Shahzad, H. F., Mehmood, A., and Choi, G. S. (2021). Diagnosis of vertebral column pathologies using concatenated resampling with machine learning algorithms. *PeerJ Comput. Sci.*, 7(e547):e547.
- Saboor, A., Usman, M., Ali, S., Samad, A., Abrar, M. F., and Ullah, N. (2022). A method for improving prediction of human heart disease using machine learning algorithms. *Mob. Inf. Syst.*, 2022:1–9.
- Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517.
- Sanz, I. and Duarte, O. (2019). Graph-based feature enrichment for online intrusion detection in virtual networks. In *Anais Estendidos do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 129–136, Porto Alegre, RS, Brasil. SBC.
- Sharma, D. K. and Hota, H. S. (2013). Data mining techniques for prediction of different categories of dermatology diseases. *Journal of Management Information and Decision Sciences*, 16:103.
- Srinivasan, S., Hyman, J. D., O’Malley, D., Karra, S., Viswanathan, H. S., and Srinivasan, G. (2020). Chapter three - machine learning techniques for fractured media. In Moseley, B. and Krischer, L., editors, *Machine Learning in Geosciences*, volume 61 of *Advances in Geophysics*, pages 109–150. Elsevier.
- Sulc, Z. and Řezankova, H. (2014). Evaluation of recent similarity measures for categorical data.
- Tan, P.-N., Steinbach, M., Karpatne, A., and Kumar, V. (2019). *Introduction to Data Mining (2nd Edition)*. Pearson Education, 2nd edition.
- Uzut, G. and Buyrukoglu, S. (2020). Hyperparameter optimization of data mining algorithms on car evaluation dataset. *Euroasia Journal of Mathematics Engineering Natural and Medical Sciences*, 7:70–76.
- Zaki, N., Mohamed, E., and Habuza, T. (2021). From tabulated data to knowledge graph: A novel way of improving the performance of the classification models in the health-care data. *SSRN Electronic Journal*.