Predição de citações de artigos da área de Bibliometria

Ismael Mendes dos Santos Junior[1][0000−0001−9412−6023] and Gustavo Medeiros de Araújo[2][0000−0003−0572−6997]

[1] Universidade Federal de Santa Catarina ismael.mendes@gmail.com
https://pgcin.ufsc.br/
[2] Universidade Federal de Santa Catarina gustavo.araujo@ufsc.br
https://pgcin.ufsc.br/

Abstract. The citation count metric is an important performance indicator for researchers and journals. This article aims to determine if the citation count of articles in the Bibliometrics area can be predicted by regression and classification machine learning algorithms. For that, 3100 articles indexed in the Web Of Science published between 2010 and 2021 were analyzed from nine quantitative independent variables. It was found that it is possible to predict the future citations of an article by means of the selected variables using machine learning algorithms.

Keywords: Machine Learning · Citation Count · Bibliometrics.

1 Descrição do problema

A quantidade de citações (citation count) ainda é uma métrica importante para avaliação do impacto de artigos científicos. A métrica, introduzida por Gross e Gross [6], parte do pressuposto de que artigos com maior impacto recebem mais citações, embora hoje se entenda que ter mais citações não signifique maior qualidade de um artigo citado. Além disso, a métrica também não serve para comparar artigos de áreas diferentes.

A métrica tem sido a base para o desenvolvimento de várias outras métricas que procuram medir a performance das publicações dos cientistas, como o Fator de Impacto (FI) [5] e o h-index [7], entre outros. Existem evidências de que a maior parte dos artigos são menos citados que os indicadores de impacto dos periódicos onde são publicados [1]. O fator de impacto, assim como essas outras métricas, mede a popularidade do periódico, não a qualidade individual de um artigo.

Na Ciência da Informação, a contagem de citações se insere numa área se dedica ao estudo da produção científica: a Bibliometria. Diversos estudos tem sido empreendidos na área no sentido de analisar as citações e descrever seu comportamento ao longo do tempo, mas poucos tem se orientado à predição das citações [4]. Nessa direção, estudos tem se utilizado do aprendizado de máquina (machine learning) para alcançar níveis satisfatórios de predição por meio de variados algoritmos.

Robson e Mousquès [9] empreenderam um estudo buscando compreender se é possível prever o total de citações nos artigos da área de modelagem ambiental a partir de variáveis bibliográficas e categóricas. Para isso, analisaram 7602 papers publicados desde 2005 para determinar o número de citações que tiveram até setembro de 2014. Eles usaram o algoritmo random forest a partir de indicadores quantitativos: o número de citações foi a variável dependente e as variáveis independentes foram o ano, a quantidade de páginas, a quantidade de autores, a posição do autor em ordem alfabética, o periódico, a quantidade de palavras do resumo e do título e se o artigo foi publicado e uma edição especial. O modelo permitiu um poder de explicação relativamente baixo (r2 de aproximadamente 29%).

Em proposta semelhante, Fu e Aliferis [4] buscaram prever de forma acurada a contagem de citações de publicações na área de biomedicina em um horizonte de 10 anos usando somente informações disponíveis no momento da publicação. As bases de dados para treinamento dos modelos de predição utilizaram dois tipos de entradas (features): variáveis associadas ao conteúdo, como título e abstract, e variáveis bibliométricas que mediram a qualidade dos autores, periódicos e instituições. Os modelos analisaram 3788 artigos e 20005 features e se tratam de modelos preditivos binários: artigo excede T citações em 10 anos desde que foi publicado, em que T = 20, 50, 100 e 500 citações. Eles utilizam o algoritmo Support Vector Machines (SVM) com o kernel heterogêneo polinominal e escolheram as features mais relevantes utilizando Markov Blanket pelo algoritmo HITON-PC. Eles conseguiram atingir um bom nível de predição, com acurácia entre 0,86 a 0,92, sendo que o modelo para o intervalo T = 500 foi o que alcançou maior capacidade de predição.

Outro problema semelhante é estudado por Chakraborty et. al. [2]: predizer a contagem de citações futuras de um artigo científico depois de um dado intervalo de tempo de sua publicação. Eles analisaram 1.5 milhões de papers da área de computação. Descobriram seis grandes categorias de padrões de contagem de citações e então propuseram um modelo estratificado de aprendizado de máquina em dois estágios: 1) primeiro mapearam o paper em uma das seis categorias utilizando o algoritmo SVM; 2) um módulo de regressão utilizando o algoritmo Support Vector Regression (SVR) é executado para aquela subcategoria para predizer a contagem de citações para aquele paper. Com isso o resultado foi 50% maior do que em um estágio apenas.

Luo et al. [8], em outra abordagem ao problema, procuram predizer citações a artigos por meio de uma análise do impacto do artigo no conhecimento já existente. Para isso usaram a abordagem de aprendizado pela representação de rede utilizando o algorítimo struc2vec para representar a complexidade dos features da rede e predizer as citações. Os autores utilizaram 1343 artigos sobre autofagia celular de 1973 até 2005 do banco de dados core da WoS. depois que montaram a rede com o struc2vec, usaram o algorítmo t-SNE para visualizar os resultados. Para agrupá-los em clusters, usaram o algorítmo k-means e gráficos boxplot para contar e mostrar a distribuição das citações em cada cluster.

Considerando a relevância do problema para a área de Bibliometria, empreendeu-se uma pesquisa sobre a temática nas bases de dados da WoS, Base de Dados em Ciência da Informação (BRAPCI) e nos mecanismos de busca Google Acadêmico e Microsoft Academic. Não foram encontrados estudos para previsão de citações de trabalhos publicados na área de Ciência da Informação, nem na subárea de Bibliometria, então presume-se, dada a abrangência da cobertura dessas bases, que sejam poucos.

Assim, a questão que norteia a pesquisa é: É possível prever a quantidade de citações de artigos da área de Bibliometria?

2 Metodologia

Os artigos analisados foram obtidos a partir da busca de artigos de periódicos na Web Of Science (WoS) pelo tópico "bibliometr*" nas categoria "INFORMATION SCIENCE LIBRARY SCIENCE", publicados entre 2010 e 2021, indexados nos índices SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI. A busca, realizada em fevereiro de 2021, retornou 3.154 artigos. Os metadados dos resultados da busca foram exportados em arquivos CSV por meio de recurso da WoS.

Para a escrita dos scripts foram utilizadas as bibliotecas numpy, matplotlib e pandas da linguagem Python via Google Colab. Os arquivos csv foram agrupados em uma planilha eletrônica do Excel e importados na plataforma. Entre os 67 campos disponibilizados pela WoS, foram selecionados os seguintes: i) Nome completo dos autores (Author Full Names); ii) Título do artigo (Article Title); iii) Nome do periódico (Source Title); iv) Quantidade de referências citadas (Cited Reference Count); v) Quantidade de citações (Times Cited, All Databases); vi) Ano de publicação (Publication Year); vii) Edição especial (Special Issue); viii) Número de páginas (Number of Pages).

Algumas variáveis utilizadas no estudo foram obtidas a partir do tratamento desses campos selecionados. A variável dependente para a análise de regressão, a contagem de citações (citation count), foi obtida pelo pré-processamento do quantidade de citações do artigo que foram convertidas para a quantidade média de citações do artigo desde a data de sua publicação. Se publicado em 2020 ou 2021 considerou-se o mesmo quantitativo de citações que o campo "Times Cited, All Databases" traz na base da WoS, visto que a coleta dos dados ocorreu no dia 17/02/2021, ainda no início do ano.

Para as análises por meio de algoritmos de classificação foi necessário estratificar as citações, visto que valores contínuos não são adequados a esses modelos. Assim, a variável modelo foi o quartil da quantidade média de citações, assumindo os valores de 1-4 correspondentes ao primeiro quartil (Q1), segundo quartil (Q2), terceiro quartil (Q3) e quarto quartil (Q4).

O campo "Author Full Names", que traz o nome completo dos autores, foi utilizado para obter a contagem de autores (Number of authors) e a posição do primeiro autor no dataset quando ordenado alfabeticamente (First Author Position). Há evidências de que a posição alfabética do primeiro nome do autor

afeta sua contagem de citações [10]. O título do artigo foi tratado para obter o número de palavras no título (Number of title words) e o mesmo foi feito quanto ao abstract (Number of abstract words). também existem evidências de que resumos mais longos estão associados a mais citações [3]. Tratamento semelhante ao dos autores foi dado ao periódico: os periódicos foram ordenados e o índice atribuído a cada um foi armazenado na variável nominal que registra qual é o periódico (Journal), de forma a se ter uma variável nominal para identificá-lo. As variáveis selecionadas foram as mesmas utilizadas no estudo de Robson e Mousquès [9]. A tabela 1 resume as variáveis utilizadas no estudo.

Table 1. Variáveis utilizadas no modelo

Variável	Descrição
Citation count	Variável dependente: contagem de citações até a data de extração
Quartile	Variável dependente: quartil da contagem de citações
Number of authors	Contagem de autores
First Author Position	Posição do primeiro autor no dataset quando ordenado alfabeticamente
Number of title words	Número de palavras no título
Number of abstract words	Número de palavras no título
Journal	Posição do periódico no dataset quando ordenado alfabeticamente
Cited Reference Count	Quantidade de citações nas referências
Publication Year	Ano da publicação
Special Issue	Se é parte de número especial
Number of Pages	Contagem de páginas

Uma vez tratados, os dados foram submetidos uma análise exploratória. Após a análise, o dataset foi submetido aos modelos preditivos. Nesse estudo foram testados os modelos de predição por regressão linear multivariada e por regressão polinomial, avaliados e comparados quanto à sua acurácia pelo coeficiente de determinação R-quadrado (R2). A variável dependente foi a quantidade de citações realizadas ao artigo (citation count) e as demais são as variáveis independentes. Também foram testados os modelos de predição por classificação por meio dos algoritmos de Regressão Logística e árvore e decisão (Decision Tree). Os modelos foram avaliados por meio das métricas de acurácia, precisão, recall e F-score. Os resultados para essas métricas foram resumidos em quadros e os resultados foram comparados entre os modelos.

A base de dados foi divida em bases de treinamento (80% dos registros) e teste (20% dos registros) com uma mostra aleatória. Como a data de publicação é um aspecto importante para se definir a média de citações, a variável dependente, os artigos foram ordenados em ordem decrescente de data de publicação e aqueles que não tinham a informação do ano da publicação foram excluídos do dataset (54 artigos), restando 3100 que foram usados na análise. Os arquivos

de projeto foram disponibilizados em repositório git: https://github.com/ismael-mendes/artigomlpgcin.

3 Resultados e discussão

Nesta seção são apresentados e analisados os resultados da análise exploratória, da predição por regressão linear e dos modelos de predição por classificação.

3.1 Análise exploratória

Entre os 3100 artigos, 19% não foram citados (591 artigos). Entre os demais, 21,7% obtiveram em média uma citação por ano e 11,9% tiveram duas citações anuais em média. Em média, os artigos foram citados pelo menos duas vezes e observou-se baixa dispersão na contagem das citações. No que se refere à distribuição entre os quartis, constatou-se que 25% tem em média entre 0 até 2 citações (Q1), que 50% tem entre 2 e até 12 citações (Q2), que 75% dos artigos tem de 12 a 31 citações (Q3) e que os demais tem mais de 31 citações (Q4).

Em média os artigos tem 2,7 autores e os dois artigos com maior quantidade de autores tem 19 pesquisadores. Metade dos artigos tem até dois autores. No que se refere ao conteúdos dos artigos, constatou-se que em média tem 16 páginas, sendo que o menor artigo tem duas páginas e o maior possui 62 páginas. O título tem em média 13 palavras e abstract, tem em média 182 palavras, apresentando esperada baixa dispersão na quantidade de palavras utilizadas. Os artigos citam em média 40 obras, sendo que o artigo que mais citou faz referência a 900 trabalhos. Verificou-se ainda que 75% dos trabalhos foram publicados até 2018.

Buscou-se avaliar também eventual existência de correlação entre as variáveis. O coeficiente de correlação de Pearson mostrou-se pouco significativo para quase todas as variáveis. Apenas verificou-se correlação moderada entre o número de páginas e a quantidade de obras citadas (p = 0,496233), o que é esperado. Também procurou-se investigar a correlação por meio do coeficiente de correlação de postos de Spearman, que também mostrou-se pouco significativo em quase todas as variáveis. Nesse caso, a correlação entre o número de páginas e a quantidade de obras citadas também foi moderada (rô = 0,502521).

3.2 Predição por modelos de regressão

A regressão linear multivariada foi aplicada por meio da biblioteca sklearn aos artigos do dataset utilizando 80% deles como base de treinamento, com uma amostra randômica. O modelo obteve um R2 = -0,00082, utilizando todas as variáveis, cujos coeficientes são apresentados na tabela 2. Embora o modelo tenha apresentado um baixo poder de explicação da variável dependente, os coeficientes e seus sinais permitiram perceber importantes relações sobre as variáveis e o total de citações. Há uma relação inversa que era esperada entre a quantidade de palavras no título e no abstract e o número de citações. Também guardam

uma relação inversa com a quantidade de citações o ano da publicação e o fato de constar em uma edição especial. As demais variáveis tem uma relação direta com a variável dependente.

Table 2. Coeficientes das variáveis pela regressão linear multivariada

Variável	Coeficiente
Number of authors	8,13833606e-02
First Author Position	4,12811714e-05
Number of title words	-1,45192502e-02
Number of abstract words	-1,36064842e-03
Journal	7,50202485e-03
Cited Reference Count	2,09894721e-02
Publication Year	-2,78107086e-01
Special Issue	-4,72825963e-01
Number of Pages	2,77739273e-02
Resíduos	560,76318713

Em seguida, a Regressão Logística também foi aplicada ao dataset. Os dados foram transformados por meio da biblioteca sklearn em funções polinomiais. Também foi utilizada a mesma amostra randômica correspondente a 80% dos artigos e foram testadas as funções de primeiro até o quinto grau. Em todas as situações o poder de explicação foi insatisfatório, sendo o R2 = -3,13938 o melhor resultado alcançado com funções polinominais de segundo grau. Constatou-se que os dois modelos se mostraram insatisfatórios para a previsão das citações, e assim partiu-se para a previsão por meio de modelos de classificação.

3.3 Predição por modelos de classificação

Os dados foram submetidos aos modelos de classificação por regressão logística e Decision Tree. Por sua natureza, modelos de classificação exigem que a variável dependente seja uma variável categórica. Assim, a variável dependente adotada foi o quartil da contagem das citações (Quartile): ao invés de predizer a contagem das citações os modelos passaram a buscar prever em qual dos quartis o artigo se situa. Por exemplo, o modelo procurará predizer se o artigo figura entre os 25% mais citados (Q1) ou entre os 75% (Q3) mais citados.

A regressão logística foi aplicada por meio da biblioteca sklearn aos artigos do dataset também utilizando 80% deles como base de treinamento com uma amostra randômica. Embora a acurácia do modelo tenha sido de 0,65968, não alcançou boa qualidade preditiva para as classes Q3 e Q4, apresentando precisão, recall e f-score melhor adequados ao primeiro quartil (Tabela 3). A precisão aponta qual foi a proporção de classificações positivas realizadas corretamente. É obtida pela relação entre verdadeiros positivos e a soma de verdadeiros positivos com falsos positivos.

O indicador recall, que é obtido pela razão entre verdadeiros positivos e a soma dos verdadeiros positivos com os falsos negativos, permite identificar

qual é a proporção de positivos identificados corretamente. Ele é maior e mais significativa para Q1 (93%) do que para Q2 (13%) como mostra a tabela 3. Quanto ao F-Score, é um indicador que relaciona precisão e recall e verifica-se que o F-Score para as classificações preditas para o Q1 foram mais satisfatórias que para as demais classes.

Table 3. Métricas do modelo de regressão logística

Variável	Precisão	Recall	F-Score
Q1	0,68	0,93	0,79
Q2	0,48	0,13	0,20
Q3	0,00	0,00	0,00
Q4	0,00	0,00	0,00

Também foi avaliada a predição da regressão por meio do kernel Radial basis function network (RBF) a partir do algoritmo Support Vector Machine (SVM) e obteve-se uma acurácia um pouco melhor, de 0,67097. Entretanto, o modelo também não se mostrou aceitável para prever as classes Q3 e Q4, apresentando precisão, recall e f-score melhores para o Q1 (Tabela 4). Observa-se que o F-Score foi um pouco menor para o Q2, o que ocorreu principalmente em função do menor valor da métrica recall para ele. Assim, reduziu-se a proporção de positivos identificados corretamente pelo modelo.

Table 4. Métricas do modelo de regressão logística com kernel SVM

Variável	Precisão	Recall	F-Score
Q1	0,68	0,96	0,80
Q2	0,53	0,10	0,17
Q3	0,00	0,00	0,00
Q4	0,00	0,00	0,00

Em seguida os dados foram submetidos ao modelo de árvore de decisão (Decision Tree). Como os demais, também utilizou-se de uma amostra aleatória correspondente a 80% do dataset para o treinamento. A matriz de confusão, que é um tabela que apresenta as frequências de classificação para cada classe do modelo, é apresentada na Tabela 5, permitindo visualizar o resultado das predições. Nota-se que os verdadeiros positivos (destacados em negrito) foram poucos para as classes Q3 e Q4.

O modelo apresentou acurácia de 0,598387 um pouco menor que a obtida com a regressão logística. Entretanto, as métricas para o modelo (Tabela 6) mostraram-se melhores para Q2 e Q3, embora tanto para o terceiro quanto para o quarto quadrante ainda sejam pouco explicados pelo modelo.

Entre os modelos de classificação, o modelo construído utilizando o algoritmo Decision Tree foi o que apresentou o melhor equilíbrio entre a acurácia e

Table 5. Matriz de confusão

	Q1	Q2	Q3	Q4
Q1	**287**	114	7	2
Q2	108	**82**	8	2
Q3	2	5	**2**	0
Q4	0	1	0	**0**

Table 6. Métricas do modelo de árvore de decisão

Variável	Precisão	Recall	F-Score
Q1	0,72	0,70	0,71
Q2	0,41	0,41	0,41
Q3	0,12	0,22	0,15
Q4	0,00	0,00	0,00

as métricas. No estudo de Robson e Mousquès [9], as variáveis quantitativas utilizadas por eles, as quais também foram utilizadas neste estudo, tiveram pouco poder de predição da frequência de citações na área de modelagem ambiental (29%). O modelo proposto por Chakraborty et. al. [2] alcançou uma acurácia de 0,78, entretanto a estratégia de estratificação adotada por eles é mais complexa do que a adotada neste estudo. De forma geral, os artigos com maior volume de citações - que são poucos - não são bem preditos pelos modelos.

4 Considerações finais

Os resultados mostram que é possível predizer a contagem de futuras citações de um artigo por meio de variáveis quantitativas obtidas a partir de pré-processamento de dados já prontamente disponíveis nas bases de dados de indexação. Os algoritmos de machine learning mostraram-se viáveis para a predição da contagem de citações, mas a abordagem utilizando a regressão linear (multivariada e polinomial) não alcançou bom poder de explicação do fenômeno. Os modelos de classificação alcançaram melhores resultados, embora também compartilhem da limitada predição de artigos com maior volume de citações.

Em estudos futuros pretende-se utilizar nos modelos as variáveis relacionadas ao conteúdo do artigo, como o abstract e palavras-chave, por meio de Neural Language Processing (NLP). Percebeu-se a oportunidade de se processar o conteúdo dos artigos para além da contagem de palavras de forma a aprimorar a qualidade dos modelos de predição de contagem de citações. É oportuno também realizar a análise com uma base de dados mais ampla e representativa da produção científica em Ciência da Informação ou mesmo comparar bases de dados para perceber padrões de publicação científica.

References

1. Bozzo, A., Oitment, C., Evaniew, N., Ghert, M.: The Journal Impact Factor of Orthopaedic Journals Does not Predict Individual Paper Citation Rate. JAAOS: Global Research and Reviews **1**(2), e007 (may 2017). https://doi.org/10.5435/jaaosglobal-d-17-00007, /pmc/articles/PMC6132331/ /pmc/articles/PMC6132331/?report=abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6132331/
2. Chakraborty, T., Kumar, S., Goyal, P., Ganguly, N., Mukherjee, A.: Towards a stratified learning approach to predict future citation counts. In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries. pp. 351–360. Institute of Electrical and Electronics Engineers Inc. (dec 2014). https://doi.org/10.1109/JCDL.2014.6970190, https://ieeexplore.ieee.org/document/6970190
3. Didegah, F., Thelwall, M.: Which factors help authors produce the highest impact research? collaboration, journal and document properties. Journal of Informetrics **7**(4), 861–873 (2013). https://doi.org/https://doi.org/10.1016/j.joi.2013.08.006, https://www.sciencedirect.com/science/article/pii/S1751157713000709
4. Fu, L.D., Aliferis, C.F.: Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. Scientometrics **85**(1), 257–270 (feb 2010). https://doi.org/10.1007/s11192-010-0160-5, https://akjournals.com/view/journals/11192/85/1/article-p257.xml
5. Garfield, E.: The History and Meaning of the Journal Impact Factor. JAMA **295**(1), 90–93 (01 2006). https://doi.org/10.1001/jama.295.1.90, https://doi.org/10.1001/jama.295.1.90
6. Gross, P., Gross, E.: Text categorization with support vector machines. College libraries and chemical education. Science **46**, 423–444 (1927)
7. Hirsch, J.E.: An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences **102**(46), 16569–16572 (2005). https://doi.org/10.1073/pnas.0507655102, https://www.pnas.org/content/102/46/16569
8. Luo, Z., He, J., Qian, J., Wang, Y., Chen, J., Lu, W.: Can scientific publication's network structural features predict its citation? In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries. pp. 485–486. Institute of Electrical and Electronics Engineers Inc., New York, NY, USA (aug 2020). https://doi.org/10.1145/3383583.3398575, https://dl.acm.org/doi/10.1145/3383583.3398575
9. Robson, B.J., Mousquès, A.: Can we predict citation counts of environmental modelling papers? Fourteen bibliographic and categorical variables predict less than 30% of the variability in citation counts. Environmental Modelling and Software **75**, 94–104 (jan 2016). https://doi.org/10.1016/j.envsoft.2015.10.007, https://www.sciencedirect.com/science/article/abs/pii/S1364815215300657
10. Tregenza, T.: Darwin a better name than Wallace? [1] (1997). https://doi.org/10.1038/385480a0, https://www.nature.com/articles/385480a0