Estimando futuras colaborações com dados sobre atividades científicas

Thiago Magela Rodrigues Dias
Doutor em Modelagem Matemática e Computacional pela Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Belo Horizonte, Minas Gerais, Brasil. Professor do Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG) - Belo Horizonte, MG - Brasil.
http://lattes.cnpq.br/4687858846001290
E-mail: thiagomagela@gmail.com

Submetido em: 26/09/2020. Aprovado em: 25/11/2020. Publicado em: 28/07/2021.

RESUMO

Em uma rede de colaboração científica, uma conexão é formada quando dois ou mais cientistas publicam um trabalho em conjunto. Nesse caso, as publicações representam as arestas e os cientistas, os nós da rede. Lançando mão de conceitos de análise de redes sociais, é possível compreender melhor o relacionamento entre os nós. O trabalho em questão tem o objetivo de realizar a predição de ligações em redes de coautoria formadas pelos doutores com currículos cadastrados na Plataforma Lattes, e que tenham, como área de atuação, as Ciências da Informação. Atualmente, a Plataforma Lattes conta com 6.6 milhões de currículos de indivíduos e representa um dos conjuntos de dados curriculares mais relevantes e reconhecidos mundialmente. Diante isso, é possível compreender o comportamento da rede e acompanhar a sua evolução ao longo do tempo. Para tanto, algumas etapas precisam ser seguidas. São elas: extração dos dados, criação das redes de coautoria, definição dos atributos a serem utilizados, criação de um conjunto de dados, e por fim, emprego dos mesmos como entrada em um algoritmo de aprendizado de máquinas. Por meio dos resultados, é possível estabelecer, com precisão, a evolução da rede de colaborações científicas dos pesquisadores a nível nacional, auxiliando, assim, as agências de fomento na escolha de futuros pesquisadores de destaque.

Palavras-chave: Colaboração científica. Predição de ligações. Plataforma Lattes.

Estimating future collaborations with data on scientific activities

ABSTRACT

In a scientific collaboration network, a connection is formed when two or more scientists publish a work together. In this case, the publications represent the edges, and the scientists represent the nodes of the network. Using concepts from the analysis of social networks, it is possible to better understand the relationship between nodes. The work in question aims to make the prediction of connections in co-authorship networks formed by PhDs with curricula registered in the Lattes Platform, and whose area of activity is Information Sciences. Currently, the Lattes Platform has 6.6 million curricula of individuals and represents one of the most relevant and recognized scientific repositories worldwide. With this, it is possible to understand the behavior of the network and monitor its evolution over time. For that, some steps are necessary, they are: data extraction, creation of co-authorship networks, definition of the attributes to be used, creation of a data set, and finally, use them as input in a machine learning algorithm. Through the results it is possible to establish, with precision, the evolution of the network of scientific collaborations of the researchers at national level, thus assisting the funding agencies in the choice of future outstanding researchers.

Keywords: *Scientific collaboration. Link prediction. Lattes Plataform.*

Estimación de colaboraciones futuras con datos sobre actividades científicas

RESUMEN

En una red de colaboración científica, se forma una conexión cuando dos o más científicos publican un trabajo en conjunto. En tal caso, las publicaciones representan los bordes y los científicos representan los nodos de la red. Aprovechando los conceptos de análisis de redes sociales, es posible comprender mejor la relación entre nodos. El trabajo en cuestión tiene como objetivo hacer la predicción de las conexiones en redes de coautoría formadas por médicos con currículos registrados en la Plataforma Lattes, y cuya área de actividad son las Ciencias de la Información. Actualmente, la Plataforma Lattes tiene 6.6 millones de CV de personas y representa uno de los repositorios científicos más relevantes y reconocidos a nivel mundial. Con esto, es posible comprender el comportamiento de la red y monitorear su evolución a lo largo del tiempo. Para eso, algunos pasos son necesarios, son: extracción de datos, creación de redes de coautoría, definición de los atributos que se utilizarán, creación de un conjunto de datos y, finalmente, utilizarlos como entrada en un algoritmo de aprendizaje de máquinas. A través de los resultados es posible establecer, con precisión, la evolución de la red de colaboraciones científicas de los investigadores a nivel nacional, ayudando así a las agencias de financiación en la elección de futuros investigadores destacados.

Palavras clave: *Colaboración científica. Predicción de enlaces. Plataforma Lattes.*

INTRODUÇÃO

No final da década de 90, diversos pesquisadores dedicaram atenção aos estudos de redes. Foram realizados trabalhos sobre a área da biologia, da internet, de roteadores, entre outros (NEWMAN, 2001; NEWMAN; PARK, 2003; BARABÁSI; ALBERT, 1999). Tais investigações permitiram entender o relacionamento entre os nós, fazendo surgir, ao se estudar essas ligações, a pergunta: "como ocorre a evolução da rede ao longo do tempo?". Hasan e Zaki (2011), porém, explicam que compreender a evolução da rede como um todo é uma tarefa complexa.

Com esses conceitos em mente, Liben-Nowell e Kleinberg (2003) propuseram o problema da predição de ligações. Inicialmente, foram utilizados métodos que calculavam a similaridade entre dois nós da rede. Quanto mais parecidos, maior a chance de possuírem uma ligação entre si. A partir de então, diversos outros métodos foram propostos para melhor resolução do problema da predição de ligações (ACAR; DUNLAVY; KOLDA, 2009; ZHOU; LÜ; ZHANG, 2009; LIU *et al.*, 2011).

Atualmente, são empregados métodos probabilísticos, métodos baseados em álgebra linear e, também, métodos que transformam esse problema em um de classificação binária, dessa forma, diversos algoritmos podem ser aplicados à sua resolução. Neste artigo, tratamos a predição de ligações como um problema de classificação, assim, algoritmos da área de sistemas de recomendação são empregados na realização dos objetivos propostos.

Aplicando tais conceitos a um domínio mais específico, podemos dirigir as atenções às redes pertencentes à comunidade científica. Ao se publicar uma comunicação científica com outro cientista, uma ligação é formada pela colaboração realizada. Nessas redes, os autores representam os nós e as colaborações científicas, as arestas (MARUYAMA; DIGIAMPIETRI, 2019). Tais redes são chamadas de redes de coautoria e serão o objeto de estudo deste artigo.

Nesse contexto, a Plataforma Lattes, mantida pelo CNPQ[1], tem sido fonte de dados de diversos trabalhos que visam a analisar redes de colaboração científica, principalmente por englobar dados de grande parte da produção científica nacional.

Cañibano e Bozeman (2009) destacam que os currículos acadêmicos são fontes de informação potenciais e extremamente abrangentes, bem como foco de investigações recentes que estudam grupos de pesquisadores. Inquirições que se valem de currículos no exame de redes sociais são ainda menos frequentes, porém, deve-se considerar a gama de trabalhos sobre análise de coautoria e os efeitos das colaborações científicas na carreira do pesquisador (DIGIAMPIETRI; SANTIAGO; ALVES, 2013; LIMA *et al.*, 2013; MENA-CHALCO; CESAR-JUNIOR, 2013).

Perez-Cervantes *et al.* (2013) introduzem novas medidas para estimar a influência da colaboração em redes científicas. A abordagem é baseada na técnica de predição de *links* e avalia como a presença ou ausência de um pesquisador afeta o processo de predição na rede em exame. Para isso, os cientistas são representados por nós em uma rede de colaboração e, após a remoção de nós, o processo de predição de *links* é realizado de forma iterativa para todos os outros nós.

Já Mena-Chalco *et al.* (2014) utilizam dados dos currículos da Plataforma Lattes para identificar e caracterizar a rede de colaboração de pesquisadores brasileiros. Essa pesquisa objetiva extrair os dados de currículos cadastrados na Plataforma Lattes, identificar automaticamente a colaboração baseada em informações bibliométricas, produzindo uma rede de colaboração, e aplicar métricas baseadas em análise topológica para compreender como ocorre a interação entre os pesquisadores.

Por sua vez, Sidone, Haddad e Mena-Chalco (2016) apresentam o papel da geografia na evolução da produção e colaboração científica no Brasil entre 1992 e 2009.

Nesses estudos, foi feito uso de dados dos currículos de um milhão de pesquisadores, abrigados na Plataforma Lattes. Os autores destacam o processo de desaceleração da produção científica brasileira a partir dos últimos triênios analisados.

Atualmente, a Plataforma Lattes conta com 6.6 milhões de currículos cadastrados e representa uma das mais relevantes fontes de dados sobre atividades científicas e pesquisadores, além de ser reconhecida mundialmente (LANE, 2010). O conjunto de dados registrados nos currículos cadastrados nesse sistema de informações possui atributos como: nome, formação acadêmica, experiência profissional, projetos, publicações científicas, entre outros. O grande volume de dados presente nos currículos pode fornecer informações valiosas e, até então, desconhecidas (DIAS *et al.*, 2013).

Dessa forma, será realizada a predição de ligações em redes de coautoria, formada pelos dados de doutores presentes em currículos cadastrados na Plataforma Lattes. Com isso, será possível compreender o comportamento dessa rede e acompanhar a sua evolução ao longo do tempo. Por meio deste estudo, também será possível identificar os pesquisadores que poderão colaborar com a rede no futuro.

METODOLOGIA

Para que seja possível atingir os objetivos propostos, é essencial que se siga alguns passos. Desse modo, nesta seção, serão destacados os métodos empregados nesta pesquisa para que seja possível realizar a predição de futuras ligações em uma área específica. Para tanto, foi escolhida a grande área de Ciências Sociais Aplicadas e, posteriormente, a área de Ciência da Informação. Esse conjunto de dados possui 1.084 pesquisadores com título de doutor. Inicialmente, será apresentado o *framework* empregado na extração dos dados. Em um segundo momento, serão mostradas as redes de colaboração científicas criadas, e por último, os atributos selecionados para a predição serão caracterizados.

[1] Conselho Nacional de Desenvolvimento Científico e Tecnológico

Para início do desenvolvimento do trabalho, foi necessário extrair os dados a serem utilizados. Sendo assim, lançou-se mão de um *framework* para extração e tratamento dos dados, o *LattesDataExplorer* (DIAS, 2016). Após a coleta dos dados, ocorrida em 2019, as informações foram organizadas e, posteriormente, as redes foram caracterizadas, conforme método para identificação de colaborações científicas em grandes bases de dados, com uso de baixo poder computacional, apresentado por Dias e Moita (2015).

Após a caracterização das redes de colaboração, foi preciso identificar os atributos utilizados na predição. Assim, um conjunto básico de características, oriundo de outros trabalhos que abordaram esse tema, foram selecionados. Ainda tendo a Plataforma Lattes como fonte de dados, foi possível obter informações referentes ao domínio que estava sendo analisado, visto que os autores possuem registro de algumas informações pessoais, como a cidade e o estado em que residem, e a universidade da qual fazem parte. Como tais campos podem ser empregados no auxílio à predição a ser realizada, dois tipos de atributos são definidos: os atributos topológicos e os atributos referentes ao domínio, conforme demonstrado no quadro 1.

Enquanto atributos topológicos são obtidos a partir de alguns cálculos, que utilizam como base a própria rede, atributos referentes ao domínio são extraídos e armazenados da mesma forma que o autor preencheu as informações de seu currículo. Porém, ao se lançar mão de técnicas de aprendizado de máquinas, é importante que os dados estejam padronizados para facilitar o processo de predição. Os campos "cidade", "estado" e "instituição" são considerados dados categóricos, e, por isso, devem passar por duas etapas antes de serem aplicados ao restante do processo.

Inicialmente, é necessário codificar os textos informados pelo autor em números. Por exemplo: no lugar de "Belo Horizonte", o valor 5 será armazenado; para "São Paulo", o valor 13; e assim por diante, para todas as informações categóricas. Para fazer essa codificação, é feito uso do método *Label Encoding*, algoritmo que realiza o passo a passo descrito acima para os dados selecionados. Dessa forma, todos os valores categóricos são codificados em números. Porém, após esse processo, os algoritmos a serem utilizados, podem identificar, a título de ilustração, que o valor 13, referente a São Paulo, é mais vultoso do que o valor 5, referente a Belo Horizonte, afinal, não foi especificado que esses valores representam categorias. Para evitar que isso aconteça, outro método, o *One Hot Encoder,* deve ser aplicado. Por meio dele, cada categoria é transformada em uma coluna, e, caso o valor seja referente a uma determinada coluna, é inserido o número 1, caso contrário, é inserido o zero. Assim, os dados categóricos são convertidos em uma grande matriz esparsa, composta, em sua maioria, pelo número zero.

Após a definição dos atributos, alguns passos devem ser seguidos. Em um primeiro momento, é necessário definir os períodos para treino e teste, de modo que três redes diferentes foram caracterizadas. Para a rede 1, foram definidas as publicações realizadas no período entre 1960 e 2000, que será chamado de período inicial. Já a segunda rede foi caracterizada pelo período de 2001 a 2010. Por fim, foi estabelecido o período de 2011 a 2018 para a terceira e última rede. Tais períodos compreendem a data do primeiro trabalho registrado na plataforma até o último ano anterior à coleta dos dados.

As informações referentes às redes são exibidas na tabela 1 e na figura 1, onde é possível perceber as mudanças ao longo dos períodos analisados. No princípio, nota-se a pequena quantidade de colaborações, que impacta diretamente no grau médio da rede, e na sua densidade.

Quadro 1 – Definição de atributos utilizados no processo de predição

Atributo	Descrição	Tipo
Vizinhos em Comum (VC)	De acordo com Liben-Nowell e Kleinberg (2003), a forma mais simples de realizar a predição de arestas é por intermédio da métrica Vizinhos em Comum, que pode ser entendida como a quantidade de nós em comum que dois nós específicos possuem.	Topológico
Coeficiente de Jaccard (JC)	Mede a probabilidade de que ambos, x e y, possuam um vizinho v, escolhido aleatoriamente. Hasan e Zaki (2011) explicam que, ao contrário do atributo Vizinhos em Comum, o coeficiente de Jaccard normaliza o número de vizinhos em comum.	Topológico
Adamic/Adar (AA)	Essa formulação atribui, às características mais raras, um peso maior. Podemos entendê-la como o número de propriedades compartilhadas pelos nós, dividido pelo log da frequência das características.	Topológico
Resource Allocation (RA)	Seguindo o mesmo raciocínio, a métrica Resource Allocation atribui peso na relação de dois nós, favorecendo as relações entre aqueles que possuem poucos relacionamentos.	Topológico
Preferential Attachment (PA)	A métrica Preferential Attachment foi proposta considerando apenas o tamanho das vizinhanças dos nós,. Em suma, ela estabelece que a probabilidade de um novo relacionamento com outros vértices é baseada no grau do nó em questão.	Topológico
Menor Caminho (MC)	O fato de que amigos de amigos podem criar uma ligação sugere que a distância entre os nós de uma rede pode influenciar na formação de novas ligações. Podemos entendê-la como o caminho mínimo entre dois nós.	Topológico
Colaborações em conjunto (Peso)	Dessa forma, é possível identificar colaboradores que já trabalham juntos há mais tempo e, possivelmente, possuem uma maior influência nos próximos instantes de tempo.	Topológico
Instituição	Instituição à qual o pesquisador está vinculado.	Domínio
Estado	Estado cadastrado no campo “endereço profissional”, no do currículo do pesquisador.	Domínio
Cidade	Cidade cadastrada no campo “endereço profissional”, no currículo do pesquisador.	Domínio

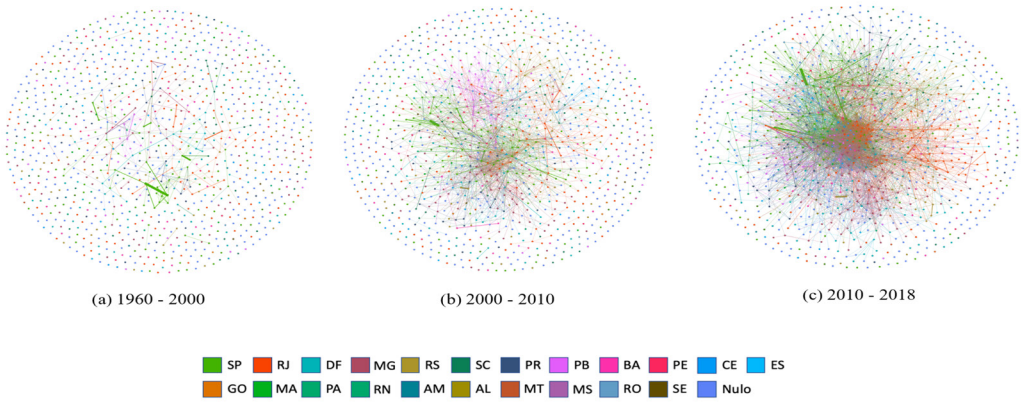
Fonte: Adaptado de Adamic, L. A. e Adar, E. (2003); Chen, H., Li, X., Huang, Z. (2005); Liben-Nowell, D. e Kleinberg, J. (2007); Potgieter, A. *et al.* (2009); Digiampietri, L. *et al.* (2015); Maruyama, W. T. e Digiampietri, L. A. (2019); Hasan, M. A. e Zaki, M. J. (2011); Lü, L. e Zhou, T. (2011).

Tabela 1 – Redes de colaboração científica caracterizadas para o estudo

Período	Pesquisadores	Colaborações	Grau médio	Densidade	Diâmetro	Caminho médio
1960 – 2000	1.084	1.064	0,466	0	15	5,523
2000 – 2010	1.084	6.186	3,701	0,003	10	4,081
2010 – 2018	1.084	11.603	11,051	0,01	8	3,280

Fonte: Elaboração dos autores.

Figura 1 – Visão geral das redes de colaboração científica caracterizadas.



Fonte: Dados da Pesquisa.

É importante examinar o aumento da densidade no decurso do tempo, uma vez que esse fator influencia diretamente no problema chamado desbalanceamento de classes, que será apresentado posteriormente. Pode-se entender o grau médio, nesse caso, como a média das colaborações realizadas pelos pesquisadores. Ao final do período analisado, o número médio de colaborações aumentou consideravelmente, refletindo, assim, na densidade da rede. Outro fator que indica a evolução da rede é a diminuição no caminho médio, demonstrando que os pesquisadores criaram mais conexões entre si, ficando, dessa maneira, mais fácil de se conectar a novos parceiros de pesquisa. Também cabe destacar que, para os propósitos desta investigação, o número de nós se manteve constante durante todos os períodos.

O conjunto de dados contendo os pesquisadores, as ligações entre si e os atributos selecionados foi então utilizado como entrada em um algoritmo de aprendizado de máquina.

Cada linha do conjunto de dados é composta pelos seguintes itens: identificação do primeiro pesquisador, identificação do segundo pesquisador, vizinhos em comum, coeficiente de Jaccard, Adamic/Adar, *Resource Allocation*, *Preferential Attachment*, Menor Caminho, peso, presença (ou ausência) de uma aresta, instituição, cidade e estado.

Nessa etapa do trabalho, o problema do desbalanceamento de classes vem à tona. O número de ligações possíveis em um grafo é quadraticamente relacionado ao número de nós, no entanto, o cômputo de ligações reais representa apenas uma pequena fração desse número (HASAN; ZAKI, 2011).

Uma técnica tradicional para superar o desbalanceamento das classes é chamada de sob amostragem. Ela consiste em reduzir o número de amostras da classe determinante, de forma randômica, igualando, assim, o número de componentes para ambos os casos. Essa técnica foi aplicada à investigação aqui apresentada.

No início, o conjunto de dados apresentava uma proporção de 152 arestas ausentes para cada aresta presente. Após a aplicação da sob amostragem, o número de arestas presentes e ausentes foi o mesmo. Com os dados balanceados, o algoritmo para predição de ligações foi executado.

RESULTADOS

Ao longo do processo descrito na seção anterior, o conjunto de dados sofreu algumas alterações. No total, os 1.084 pesquisadores podem possuir um total de 586.896 arestas. Destas, apenas 3.831 representavam arestas positivas na Rede 3. Logo, por meio do balanceamento das amostras, um conjunto randômico de outras 3.831 arestas ausentes foi escolhido. Sendo assim, o conjunto de dados utilizado na entrada no algoritmo de predição de dados é composto por 7.662 registros. Ao se fazer uso dos métodos de aprendizado de máquinas, é importante separar uma parte do conjunto de informações para treinar o algoritmo, e outra parte para o teste do mesmo. Esse segundo conjunto deve possuir dados até então não empregados em algum momento pelos algoritmos de predição, de modo a validar que realmente ocorreu um aprendizado, e não apenas um condicionamento dos valores já utilizados. Dessa forma, foram selecionadas 5.746 ligações (escolhidas aleatoriamente) para treino, representando 25% do conjunto total, e outras 1.916 ligações para teste.

Diversos algoritmos podem ser aplicados à resolução de problemas de classificação. Entre eles, alguns foram selecionados para execução do trabalho, tais como: Regressão Logística, K-Vizinhos Mais Próximos, Baías Ingênuas e Florestas Aleatórias.

Cada uma dessas técnicas possui uma particularidade diferente e, por conseguinte, consequências distintas. Portanto, seus resultados serão evidenciados na tabela 2, lançando mão das métricas: precisão, revocação, F1 e área sob a curva (AUC). Como, normalmente, a maioria dos autores faz uso da área sob a curva em algoritmos empregados na predição de ligações, ela também é utilizada como base nesta análise.

Tabela 2 – Resultados utilizando atributos topológicos da rede

Algoritmo	Precisão	Revocação	F1	AUC
Regressão Logística	0.67	0.66	0.65	0.70
K-Vizinhos Mais Próximos	0.71	0.68	0.68	0.71
Baías Ingênuas	0.76	0.62	0.56	0.70
Florestas Aleatórias	0.70	0.68	0.67	0.71

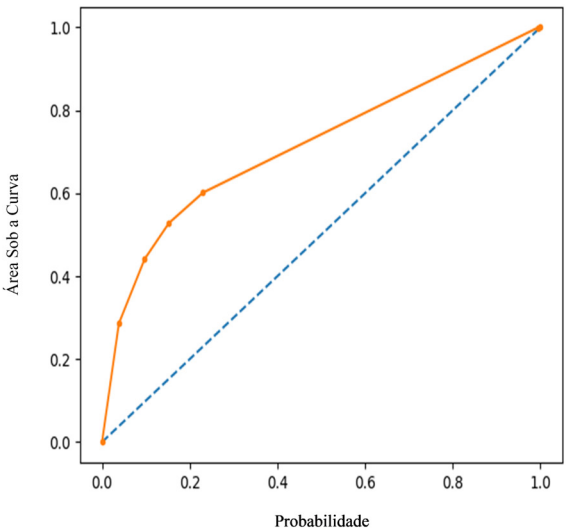
Fonte: Dados da Pesquisa, 2019.

Cada uma das métricas aplicadas à validação dos resultados possui características próprias. A precisão tem como objetivo responder à seguinte pergunta:De todos os valores preditos positivos, quantos realmente estão corretos? Uma alta precisão está relacionada a poucos falsos positivos. Considerando todos os valores positivos, a revocação tem o objetivo de saber quantos destes foram realmente preditos. A métrica F1 leva em conta a precisão e a revocação, fazendo, assim, uma média ponderada dessas duas métricas. Por último, a área sob a curva, ou *area under the curve* (AUC), em inglês, é utilizada para exibir o desempenho de um modelo de classificação no decurso de todo o processo de aprendizagem.

Foram realizados dois processos de predição: o primeiro deles, servindo-se apenas dos atributos topológicos da rede; e, posteriormente, de todo o conjunto de dados, contendo os dois tipos de atributos: topológicos e relacionados ao domínio. Dessa forma, também, é possível examinar a importância de se estudar o contexto no qual a predição de ligações será realizada.

Analisando a tabela 2, que comporta os dados referentes ao processo de predição, fazendo uso apenas dos atributos topológicos, é possível perceber que os algoritmos escolhidos obtiveram bons resultados. Dessa forma, fica claro que o algoritmo conseguiu empregar o conjunto de dados e características ora apresentado para realizar predições corretas a respeito de futuras ligações. Ao observar a área sob a curva, percebemos que todos obtiveram um resultado acima do que um mero acaso. Essa situação é visualizada com mais nitidez na figura 2, onde a linha pontilhada em azul representa uma chance de 50% de acerto, ou seja, probabilidades iguais para a predição ser da classe correta ou incorreta, e a linha laranja representa os valores das predições realizadas.

Figura 2 – Área sob a curva (AUC) para o algoritmo K-Vizinhos Mais Próximos.



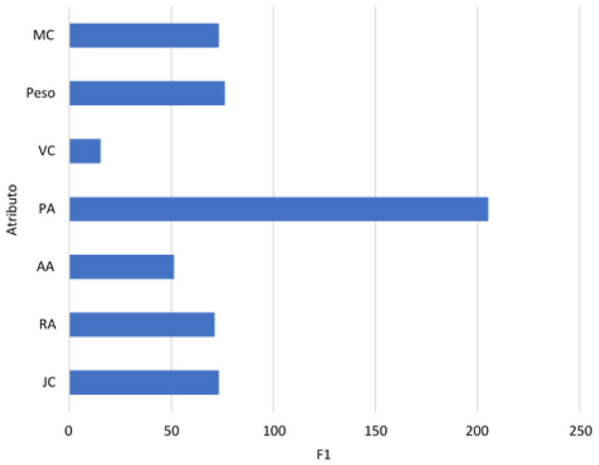
Fonte: Dados da Pesquisa, 2019.

Entre os algoritmos utilizados, o que demonstrou melhor desempenho, levando em conta todas as métricas, foi o K-Vizinhos Mais Próximos, seguido por Florestas Aleatórias, Baías Ingênuas, e, por último, Regressão Logística. Porém, existe uma pequena diferença entre os resultados obtidos, deixando claro que, para o problema em questão, ainda não podemos estabelecer qual técnica deveria ser tomada como padrão.

Ao se analisar o processo de aprendizado com base nos atributos utilizados, é possível identificar a ordem de influência de cada um deles no resultado final. Podemos observar, a partir da figura 3, que a ordem de importância dos atributos para a predição aqui realizada é: *Preferential Attachment*, Peso das Colaborações, Caminho Mais Curto, Coeficiente de Jaccard, *Resource Allocation*, Adamic/Adar, e, por fim, Vizinhos em Comum.

Tal fato apresenta um comportamento até então diferente da maioria dos referenciais teóricos aqui estudados, em que, na maior parte das vezes, o atributo mais relevante é o Vizinhos em Comum. Já, nos estudos aqui realizados, a métrica *Preferential Attachment* é responsável por boa parte do resultado final.

Figura 3 − Ordem de importância dos atributos para a predição



Fonte: Dados da Pesquisa, 2019.

Tabela 3 − Resultados considerando todos os atributos da rede

Algaritmo	Precisão	Revogação	F1	AUC
Regressão Logística	0.78	0.78	0.78	0.87
K-Vizinhos Mais Próximos	0.74	0.73	0.73	0.80
Baías Ingênuas	0.77	0.63	0.63	0.63
Florestas Aleatórias	0.77	0.77	0.77	0.85

Fonte: Dados da Pesquisa, 2019.

Ao olhar com atenção para toda a base de dados, contendo todos os atributos, os resultados revelaram uma expressiva melhora, conforme tabela 3.

Em média, as predições foram 8,25% melhores, sendo, a melhor técnica, a Regressão Logística, que conseguiu identificar futuras colaborações com 87% de certeza. Em segundo lugar, o algoritmo Florestas Aleatórias apresentou um resultado bem próximo, com 85% de acerto, seguido pelos K-Vizinhos Mais Próximos, com 80% de acerto, e, finalmente, as Baías Ingênuas.

Nesse segundo momento, os atributos não foram analisados separadamente, visto que os dados categóricos foram codificados em diversas colunas, transformando, assim, o conjunto de dados em uma matriz esparsa. Todas as outras métricas também apresentaram um resultado melhor do que apenas utilizando os atributos topológicos da rede.

CONSIDERAÇÕES FINAIS

Os resultados aqui expostos demonstram que é possível realizar a predição de ligações lançando mão de informações da própria rede estudada. O objetivo proposto foi então alcançado, uma vez que, a partir da utilização desses dados, é possível saber, por exemplo, se dois pesquisadores da área citada acima irão colaborar em um futuro instante de tempo. Um dos pontos mais importantes desta investigação está relacionado com a evolução da rede de colaboração científica. Com o passar do tempo, as colaborações saíram de uma média de 0,46 para 11,05 por pesquisador, demonstrando que o trabalho em equipe é cada vez mais necessário.

Observando os resultados obtidos com base nas predições, fica clara a importância de se possuir conhecimento sobre o domínio a ser analisado. Inicialmente, fazendo uso apenas de atributos topológicos, ou seja, referentes à própria rede, a melhor taxa de acertos foi de 71%, para os algoritmos K-Vizinhos Mais Próximos e Florestas Aleatórias, onde o atributo mais relevante foi o *Preferential Attachment*. Esse atributo demonstra que a probabilidade de um autor publicar uma comunicação científica varia conforme o número de colaborações já realizadas.

Considerando também os atributos referentes ao domínio, nesse caso, a instituição, a cidade e o estado do pesquisador, o aumento na quantidade de predições corretas foi expressivo, saltando de 70,5%, em média, para 78,75%. Desse modo, o algoritmo com o melhor resultado foi a Regressão Logística, que realizou a predição correta em 86% dos casos.

Os atributos categóricos utilizados para que houvesse essa melhora passaram por um longo processo de codificação, a fim de que tais resultados fossem alcançados. Dessa forma, é evidente o mérito do emprego das técnicas mais avançadas de aprendizado de máquinas, para que seja possível aumentar o número de predições corretas.

Em trabalhos futuros, destaca-se a importância de aumentar o conjunto de dados ou, até mesmo, buscar outras formas de solucionar o problema do desbalanceamento de classes, aumentando, assim, o número de amostras presentes para treino do algoritmo. Desse ponto em diante, espera-se que os classificadores apresentem um desempenho ainda melhor.

REFERÊNCIAS

ACAR, E.; DUNLAVY, D. M.; KOLDA, T. G. Link prediction on evolving data using matrix and tensor factorizations. In: *Proceedings of the workshop on large-scale data mining:* theory and applications (LDMTA'09). [s.l]: [s.n], 2009. p. 262-269.

ADAMIC, L. A.; ADAR, E. Friends and neighbors on the web. *Social Networks, Elsevier*, v. 25, n. 3, p. 211-230, 2003.

BARABÁSI, A. L. E.; ALBERT, R. Emergence of scaling in random networks. *Science, American Association for the Advancement of Science*, v. 286, n. 5439, p. 509-512, 1999.

CAÑIBANO, C.; BOZEMAN, B. Curriculum vitae method in science policy and research evaluation: the state-of-the-art. *Research Evaluation*, v. 18, n. 2, p. 86-94, 2009.

CHEN, H.; LI, X.; HUANG, Z. Link prediction approach to collaborative filtering. *Proceedings Of The ACM/IEE Joint Conference on Digital Libraries*, p. 141-142, 2005.

DIAS, T. M. *et al.* Modelagem e caracterização de redes científicas: um estudo sobre a Plataforma Lattes. Brasnam-Ii Brazilian Workshop On Social Network Analysis And Mining, p. 10-20, 2013.

DIAS, T. M. R.; MOITA, G. F. A method for the identification of collaboration in large scientific databases. *Em Questão*, v. 21, n. 2, p. 140-161, 2015.

DIAS, T. *Um estudo da produção científica brasileira a partir de dados da Plataforma Lattes*. Tese (Doutorado em Modelagem Matemática e Computacional) - Centro Federal de Educação Tecnológica de Minas Gerais. Belo Horizonte, 181 p., 2016.

DIGIAMPIETRI, L. A; SANTIAGO, C. R. N.; ALVES, C. M. Predição de coautorias em redes sociais acadêmicas: um estudo exploratório em Ciência da Computação. In: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING, 2, 2013, *Anais…* Maceió, 2013.

DIGIAMPIETRI, L. *et al.* Um sistema de predição de relacionamentos em redes sociais. *Brazilian Symposium on Information Systems*, v. 11, 2015.

HASAN, M. A.; ZAKI, M. J. A survey of link prediction in social networks. In: AGGARWAL, C. (Ed.). *Social network data analytics*. Boston: Springer, 2011, p. 243–275.

LANE, J. Let's Make Science Metrics More Scientific. *Nature*, v. 464, n. 7288, p. 488-489, 2010.

LIBEN-NOWELL, D.; KLEINBERG, J. The link-prediction problem for social networks. *Journal of The American Society For Information Science And Technology*, v. 58, n. 7, p. 1019-1031, 2007.

LIMA, H. *et al.* Aggregating productivity indices for ranking researchers across multiple areas. In: Proceedings of the 13th acm/ieee-cs joint conference on digital libraries, ACM, p. 97-106, 2013.

LIU, Z. *et al.* Link prediction in complex networks: a local naïve bayes model. *EPL (Europhysics Letters)*, v. 96, n. 4, 2011.

LÜ, L.; ZHOU, T. Link prediction in complex networks: a survey. *Elsevier*, v. 390, n. 6, p. 1150-1170.

MARUYAMA, W. T.; DIGIAMPIETRI, L. A. Co-Authorship Prediction In Academic Social Network. In: Workshop Brasileiro de Análise de Redes Sociais e Mineiração, V., 2019, Porto Alegre. *Anais…* Porto Alegre: Sociedade Brasileira de Computação, [2016].

MENA-CHALCO, J. P.; CESAR-JUNIOR, R. M. Prospecção de dados acadêmicos de currículos Lattes através de scriptLattes. In: HAYASHI, M. C. P. I.; LETA, H. E. J. (Orgs.). *Bibliometria e Cientometria*: reflexões teóricas e interfaces. São Carlos: Pedro & João, p. 109-128, 2013.

MENA-CHALCO, J. P.; *et al.* Brazilian bibliometric coauthorship networks. *Journal of the Association for Information Science and Technology*, v. 65, n. 7, p. 1424-1445, 2014.

NEWMAN, M. E. The structure of scientific collaboration networks. *Proceedings of the national academy of sciences,* v. 98, n. 2, p. 404-409, 2001.

NEWMAN, M. E.; PARK, J. Why social networks are different from other types of networks. *Physical Review E.*, v. 68, n. 3, 2003.

PEREZ-CERVANTES, E. *et al.* Using link prediction to estimate the collaborative influence of researchers. In: IEEE 9TH INTERNATIONAL CONFERENCE ON ESCIENCE (ESCIENCE), IX. *Anais…* China, Beijin, p. 293-300, 2013.

POTGIETER, A. *et al.* Temporality in Link Prediction: Understanding Social Complexity. *Emergence, Complexity & Organization*, v.11, n.1, p.69-83, 2009.

SIDONE, O. J. G.; HADDAD, E. A.; MENA-CHALCO, J. P. A Ciência nas Regiões Brasileiras: Evolução da Produção e das Redes de Colaboração Científica. *Transinformação,* v. 28, n. 1, p. 15-31, 2016.

ZHOU, T., LÜ, L., ZHANG, Y.-C. Predicting Missing Links Via Local Information. *The European Physical Journal*, v.71, n.4, p. 623–630, 2009.