XL CILAMCE
IBERO-LATIN AMERICAN CONGRESS ON COMPUTATIONAL METHODS IN ENGINEERING
NOVEMBER 11-14, 2019
Praiamar Natal Hotel & Convention
Natal, RN-BRAZIL

Prediction Brazilian's Information Science Future Co-Authorships

Felipe Affonso
Thiago Magela Rodrigues Dias
felipe-affonso@hotmail.com
thiagomagela@gmail.com
Affiliation
Av. Amazonas, 7675 - Nova Gameleira, Belo Horizonte - MG, 30510-000y

Abstract. When publishing a work together with other scientists, a connection is formed by the collaboration done. Papers written together represent the edges, and the authors represent the nodes of the network. By using the concepts of social network analysis, it is possible to better understand the relationship between these nodes. At this point, the following question arises: "How does the evolution of the network occur over time?". In order to answer this question, it is necessary to understand how two nodes interact with one another, that is, what factors are essential for a new connection to be created. The purpose of this paper is to predict connections in co-authorship networks formed by doctors with resumes registered in the Lattes Platform in the area of Information Sciences. Currently, the Lattes Platform has 6.1 million resumes from researchers and represents one of the most relevant and recognized scientific repositories worldwide. With this, it will be possible to understand the behavior of the network and monitor its evolution over time. Through this study, it will also be possible to identify the researchers who can collaborate in a future moment of time. To this end, recommendation systems will be used, these represent a specific approach to the concepts of machine learning. Through the use of this technique it is possible to understand which attributes of the nodes make them closer to each other, and therefore have a greater chance of creating a connection between them in the future. This work is extremely relevant because it uses a data set that has been little used in previous studies. Through the results it will be possible to establish the evolution of the network of scientific collaborations of researchers at national level, thus helping the development agencies in the selection of future outstanding researchers.

Keywords: Networks, Scientific Collaboration, Lattes Plataform

1 Introdução

No final da década de 90, diversos pesquisadores dedicaram atenção aos estudos de redes. Foram realizados trabalhos sobre a área da biologia, a internet, roteadores, entre outros [1–3]. A partir deste momento, as redes sociais se tornaram o foco das pesquisas. Também foram realizados trabalhos em diversos tipos de redes, com o objetivo de entender suas propriedades e características [4]. Baseado nisso foi possível representá-las matematicamente, o que impulsionou ainda mais o avanço dos trabalhos que objetivaram analisar as redes caracterizadas. Para tanto foram adotadas métricas, teorias e índices para medir o comportamento das redes. Também foram realizados trabalhos para diferenciar redes sociais de redes não sociais [2].

A partir da análise de redes, é possível explicar diversos fenômenos. A analise das redes sociais permite entender o relacionamento entre os nós. Ao se estudar essas ligações entre os nós por algum tempo, surge a pergunta "como ocorre a evolução da rede ao longo do tempo?", porém Al Hasan e Zaki [5] explica que compreender a evolução da rede como um todo é uma tarefa complexa.

Com esses conceitos em mente, Liben-Nowell e Kleinberg [6] propuseram o problema da predição de ligações. Inicialmente foram utilizados métodos que calculavam a similaridade entre dois nós da rede. Quanto mais parecidos os nós, maior a chance de possuírem uma ligação entre si.

Portanto, diversos outros métodos foram propostos para melhor resolução do problema da predição de ligações [7–9]. Atualmente, utiliza-se métodos probabilísticos, métodos baseados em álgebra linear, e também, os métodos que transformam esse problema em um de classificação binária, dessa forma, diversos algoritmos podem ser utilizados pra sua resolução. Neste trabalho, trataremos o problema da predição de ligações como um problema de classificação, dessa forma, algoritmos da área de sistemas de recomendação são utilizados para realização dos objetivos propostos.

Aplicando tais conceitos a um domínio mais especifico, podemos dirigir as atenções às redes pertencentes à comunidade cientifica. Ao se publicar um trabalho com outro cientista, uma ligação é formada pela colaboração realizada. Nestas redes os autores representam os nós, e as colaborações científicas representam as arestas [10]. Tais redes são chamadas de redes de co-autoria, e serão o objeto de estudo deste trabalho.

Neste contexto, a Plataforma Lattes, mantida pelo CNPQ[1], tem sido fonte de dados de diversos trabalhos que visam analisar redes de colaboração cientificas, principalmente por englobar dados de grande parte da produção científica nacional. Atualmente, a Plataforma Lattes conta com 4 milhões de currículos de pesquisadores e representa uma das fontes de dados científicos mais relevantes e reconhecidos mundialmente [11]. O conjunto de dados, registrados nos currículos cadastrados na Plataforma Lattes possui atributos como: nome, formação acadêmica, experiencia profissional, projetos, publicações cientificas, entre outros. O grande volume de dados presente nos currículos pode fornecer informações valiosas e até então desconhecidas [12].

Entender a evolução da rede requer compreender como dois nós interagem entre si. A rede é formada pelo relacionamento entre os nós, portanto, busca-se uma forma de prever quais pesquisadores produzirão um artigo em conjunto no futuro. Tal comportamento está presente basicamente em todas as redes sociais através das "ligações sugeridas". Dessa forma, é possível utilizar as mesmas técnicas para as redes de colaboração científica estudas neste trabalho.

Diante disso surgem os sistemas de recomendação, que representam uma abordagem especifica dos conceitos de aprendizagem de máquinas. Através do emprego dessa técnica é possível compreender quais atributos dos nós os fazem ser mais próximos uns dos outros, e, portanto possuírem uma maior chance de criarem uma ligação entre si no futuro.

Portanto, será realizada a predição de ligações em redes de coautoria formada pelos dados de doutores com currículos cadastrados na Plataforma Lattes na área de Ciência da Informação. Com isso, será possível compreender o comportamento dessa rede e acompanhar a sua evolução ao longo do tempo. Através deste estudo, também será possível identificar os pesquisadores que poderão colaborar em um futuro instante do tempo. Em um segundo momento, partindo da analise proposta, também se torna

[1] Conselho Nacional de Desenvolvimento Científico e Tecnológico

possível identificar os pesquisadores mais influentes na rede de coautorias.

O texto está organizado da seguinte forma: na Seção 2 são apresentados os trabalhos relacionados à este bem como a definição de alguns conceitos que serão importantes para compreensão do trabalho. A Seção 3 apresenta os métodos utilizados, serão explicadas todas as técnicas e decisões tomadas para conclusão do trabalho. Os resultados obtidos a partir dessa metodologia serão apresentados na Seção 4. Por fim, uma conclusão, e alguns trabalhos futuros são apresentados no Seção 5.

2 Revisão de Literatura

Em um trabalho seminal, Liben-Nowell e Kleinberg [6] propõe o problema da predição de ligações. Seu estudo é até hoje considerado o ponto de partida para este campo. O tema é introduzido com foco nas redes sociais e no seu dinamismo. Ao longo do tempo, novas arestas são adicionadas às redes, o que representa o surgimento de novas interações na estrutura social. Os autores definem o problema da predição de links como: dada uma rede social em um momento *t*, o objetivo é prever, com precisão, as arestas que serão adicionadas à rede durante o intervalo *t* e um tempo futuro *t'*. A predição de ligações, nesse contexto, permite descobrir indivíduos que já estão trabalhando juntos, porém a sua interação ainda não foi diretamente observada [13].

Com este mesmo objetivo em mente, Adamic e Adar [14] realizam um estudo a respeito de qual fonte de informação seria capaz de indicar relacionamentos entre usuários. Ao longo do trabalho são realizadas diversas etapas para entender a ligação de um usuário com outro. Nesse trabalho em questão o autor se refere ao problema como "predição de relacionamentos", e, utiliza um ranqueamento de pessoas similares para prever as arestas faltantes. Ao final do estudo, uma parcela dos estudantes recebeu a listagem das pessoas mais similares a elas, e, frequentemente, reconheciam tais indivíduos. O autor salienta que o grande desafio de tais análises é possuir apenas um pequeno conjunto de dados, que representa uma ínfima porção dos dados reais.

Porém, para que uma predição seja realizada, é necessário que conceitos relacionados às características topológicas da rede sejam melhor compreendidos. Para tanto, Newman e Park [2], realiza um trabalho que tem como foco analisar as principais diferenças entre redes sociais e não sociais. É destacado que a relação entre os graus dos nós adjacentes das redes são positivamente correlacionados nas redes sociais, porém negativamente nos outros tipos de rede. Em segundo lugar, redes sociais mostram alto nível de agrupamento. Como conclusão, redes sociais são divididas em comunidades, enquanto as não sociais, não são. Neste contexto, podemos entender os graus de uma rede como a distancia miníma, em termos de números de áreas na rede, entre todos os pares de nós na rede, pelos quais uma conexão existe [1].

Mesmo após diversos estudos na área de analise de redes sociais, Al Hasan e Zaki [5] destaca que compreender toda a evolução de uma rede é uma tarefa complexa, porém, entender a associação entre dois nós específicos é bem mais simples. Para tanto, algumas perguntas podem ser feitas: Como o padrão de associações muda com o tempo? Quais são os fatores que orientam essas associações? Como a associação entre dois nós é afetada por outros nós? Para responder as perguntas, o autor utiliza a formulação do problema dada por Liben-Nowell e Kleinberg [6] e realiza uma pesquisa das abordagens existentes com foco, principalmente, em grafos de redes sociais.

Voltando as atenções às redes de colaboração científica, objeto de estudo deste trabalho, Newman [15] apresenta um dos primeiros trabalhos a respeito deste tópico. São estudadas três redes específicas, uma de pesquisas na área biomédica, outra em física, e por último, matemática. O autor apresenta diversas características das redes de coautoria, e realiza diversas análises para compreender o comportamento dos nós nessa rede. É destacada a importância de tais redes, e como elas possuem informações meticulosas, bem documentadas e até mesmo com eventos temporais do relacionamento social e profissional dos cientistas.

Utilizando a Plataforma Lattes como fonte de dados, Dias et al. [12] descreve uma abordagem para extração dos currículos dos pesquisadores e a construção de uma rede de colaboração científica. A relação entre os colaboradores é realizada através da presença de um ou mais trabalhos em conjunto.

Através do *framework* construído, são apresentadas redes que possuem termos em comum, participaram de um mesmo congresso ou até mesmo de uma mesma área. Já em Dias et al. [16], os autores apresentam o método detalhadamente, são realizados testes para formação das redes e analisadas as propriedades presentes nas mesmas.

Perez-Cervantes et al. [17] apresenta uma abordagem que tem como objetivo obter a influência de um indivíduo em uma rede de colaboração. Para tanto, é utilizado um preditor de ligações baseado em métricas locais da estrutura da rede. A influência individual colaborativa é obtida levando em consideração a influência de um determinado pesquisador na predição de ligações da rede como um todo. Em Perez Cervantes [18] a descrição dos métodos utilizados é detalhada. São utilizados dados de 47.555 pesquisadores da Plataforma Lattes, que foram obtidos através da utilização do *ScriptLattes* [19]. Como resultado, as medidas da influência colaborativa, apresentam correlação inversa significativa quando comparadas com as medidas de centralidade mais conhecidas. Tal fato demonstra a efetividade das métricas propostas. Outro fator importante é que a metodologia descrita pode ser calculada independentemente para cada vértice, sem a necessidade de um cálculo global, reduzindo assim o custo computacional.

3 Materiais e Métodos

Para que seja possível atingir os objetivos propostos, alguns passos se fazem necessários. Nesta seção serão destacados os métodos utilizados para que seja possível realizar a predição de futuras ligações em uma área específica. Para tanto, foi escolhida a grande área de Ciências Sociais e Aplicadas e posteriormente a sub-área Ciência da Informação. Este conjunto de dados possui 1.094 pesquisadores com título de doutor. Inicialmente será apresentado o *framework* utilizado para extração dos dados. Em um segundo momento, as redes de colaboração científicas criadas, e por último, os atributos selecionados para a predição serão caracterizados.

Para início do desenvolvimento do trabalho, foi necessário realizar a extração dos dados a serem utilizados. Para tanto, o *LattesDataExplorer* [20], um *framework* para extração e tratamento dos dados foi utilizado. Como pode ser observado na Figura 1, inicialmente os dados são coletados através do CNPq e armazenados em um repositório local, onde é realizada uma seleção dos dados. Utilizando o identificador de cada currículo, a data da ultima atualização é comparada com o repositório no CNPq, caso as datas sejam divergentes, o extrator substitui o currículo que estava armazenado localmente pela versão mais atual [20]. Posteriormente os dados são tratados e armazenados em formato XML (*Extensible Markup Language*), com isso é possível gerar métricas e fazer o cálculo de algumas estatísticas.

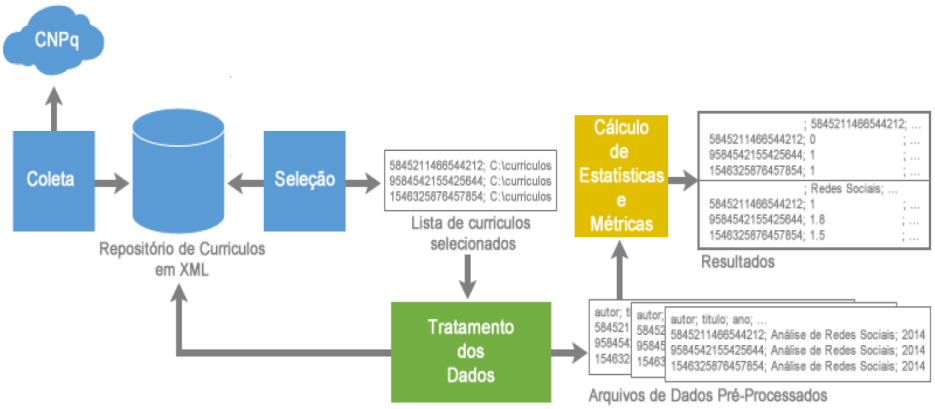


Figura 1. *Framework* utilizado para extração dos dados[20].

Já com os dados extraídos e organizados, é necessário realizar a criação das redes. De acordo com Newman [15] a coautoria de um artigo pode ser entendida como a documentação de uma colaboração

entre dois ou mais autores, e estas colaborações formam uma "rede de colaboração científica". Dias e Moita [21] apresenta um método para identificação de colaborações científicas em grandes bases de dados, com a utilização de baixo poder computacional. Portanto, este método foi utilizado para geração das redes utilizadas neste trabalho.

Após a criação das redes de colaboração, se faz necessário identificar quais atributos serão utilizados para a predição. Para tanto, um conjunto básico de características, oriundos de outros trabalhos que abordaram esse tema foram selecionados [5, 10, 14, 22–24].

De acordo com Liben-Nowell e Kleinberg [6], a forma mais simples de realizar a predição de arestas, é através da métrica vizinhos em comum. que pode ser entendida como a quantidade de nós em comum que dois nós específicos possuem. Ao utilizar esse atributo nas redes de colaboração científica, Newman [25] destaca que indivíduos que nunca trabalharam juntos, porém possuem um colaborador em comum, têm uma probabilidade muito maior de colaborar no futuro. O atributo *Common Neighbors* (CN) é demonstrado na Eq. (1), onde *x* e *y* representam vértices do grafo.

$$CN(x, y) = |\Gamma(x) \cap \Gamma(y)| \tag{1}$$

Outra métrica que pode ser obtida utilizando as características estruturais da própria rede é chamada de Coeficiente de Jaccard, e mede a probabilidade de que ambos x e y possuam um vizinho v, escolhido aleatoriamente que x ou y possuam. Al Hasan e Zaki [5] explicam que ao contrário do atributo *Common Neighbors*, o coeficiente de Jaccard normaliza o número de vizinhos em comum, conforme abaixo:

$$CoeficientedeJaccard(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \tag{2}$$

Com o objetivo de estabelecer a similaridade entre duas páginas, Adamic e Adar [14] propuseram a métrica Adamic/Adar. Para que seja possível a sua utilização em algoritmos de predição de ligações, Liben-Nowell e Kleinberg [26] customizaram a métrica conforme apresento na Eq. (3). Essa formulação atribui às características mais raras um peso maior [27]. Podemos entendê-la como o número de propriedades compartilhadas pelos nós, dividido pelo *log* da frequência das características.

$$Adamic/Adar(x, y) = \sum_{w \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\log |\Gamma(w)|} \tag{3}$$

Seguindo mesmo raciocínio, a métrica *Resource Allocation* atribui peso na relação de dois nós favorecendo as relações entre aqueles que possuem poucos relacionamentos [22], e pode ser encontrada na Eq. (4).

$$ResourceAllocation = \sum_{w \in \Gamma(u) \cap \Gamma(v)} \frac{1}{|\Gamma(w)|} \tag{4}$$

Considerando apenas o tamanho das vizinhanças dos nós, a métrica *Preferential Attachment* foi proposta, e é apresentada na Eq. (5). Em suma, estabelece que a probabilidade de um novo relacionamento com outros vértices é baseada no grau do nó em questão[5]. Essa métrica não requer informações relacionadas à vizinhança de cada nó, consequentemente possui um custo computacional mais baixo [24].

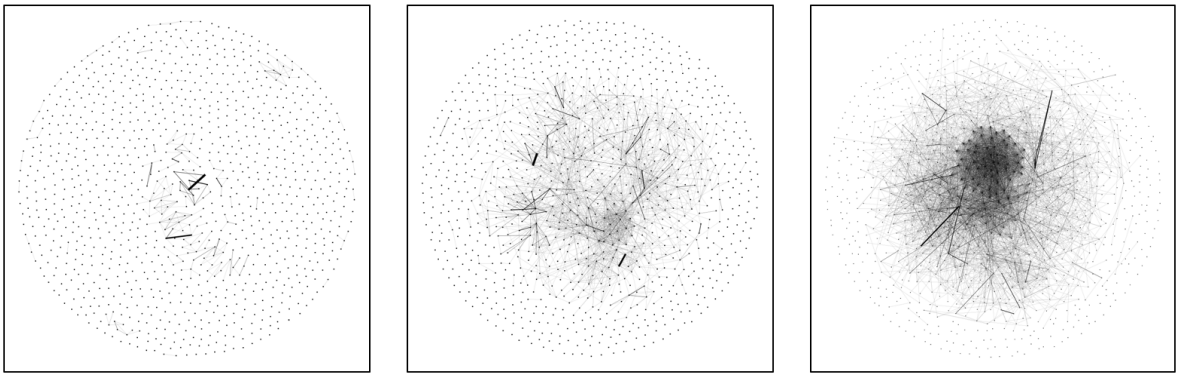
$$PreferentialAttachment = |\Gamma(u)||\Gamma(v)| \tag{5}$$

O fato de que amigos de amigos podem criar uma ligação sugere que a distância entre os nós de uma rede podem influenciar na formação de novas ligações [5]. Dessa forma, a métrica Menor Caminho também pode ser utilizada para que a predição de ligações seja realizada. Podemos entendê-la como o caminho mínimo entre dois nós [22].

Por fim, a quantidade de colaborações em conjunto que dois nós possuíram ao longo daquele instante de tempo também foi considerado como atributo a ser utilizado. Dessa forma é possível identificar colaboradores que já trabalham juntos a mais tempo, e possivelmente possuem uma maior influência nos próximos instantes de tempo.

3.1 Método Proposto

Após definir os atributos que serão utilizados, alguns passos se fazem necessários. Em primeiro momento é necessário definir os períodos para treino e teste, portanto, 3 redes diferentes foram criadas. 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 criada para o 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 finalizado anteriormente a apresentação deste trabalho. A Figura 2 apresenta as 3 redes, é possível observar que com o passar do tempo as colaborações entre os cientistas aumentou. Através dessa representação também é possível entender o objetivo do trabalho em questão. Dada a primeira rede, prever quais serão as colaborações no próximo instante de tempo.



(a) Rede 1: 1960 – 2000; (b) Rede 2: 2001 – 2010; (c) Rede 3: 2011 – 2018.

Figura 2. Evolução da Rede de Colaboração Científica da Sub-Área de Ciência da Informação

A Tabela 1 apresenta as principais características das redes criadas. A quantidade de arestas aumenta consideravelmente ao longo dos anos. Os atributos citados acima foram identificados a partir da utilização de tais redes, a maioria deles utiliza atributos topológicos para que a métrica seja calculada. Também é importante salientar que o número de pesquisadores não foi alterado ao longo do tempo, o mesmo grupo de nós foi selecionado para todo o período.

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, CN (*Common Neighbors*), Coeficiente de Jaccard, Adamic/Adar, *Resource Allocation*,*Preferential Attachment*, Menor Caminho, peso, e por fim, a presença, ou não, de uma aresta. É importante salientar que os índices correspondem aos cálculos apresentados anteriormente para os dois nós da linha. Já a aresta é obtida usando os dados do período posterior. Ou seja, dado este conjunto de atributos, uma nova aresta será gerada? Essa informação será enviada para o algoritmo de predição.

Nessa etapa do trabalho, o problema do desbalanceamento de classes vêm a tona. O número de ligações possíveis em um grafo é quadraticamente relacionado ao número de nós, no entanto, o número de ligações reais representa apenas uma pequena fração deste número [5]. De acordo com Menon e

Rede	**Período**	**Número de Nós**	**Número de Arestas**	**Grau Médio**
Rede 1	1960 - 2000	1084	191	0.3524
Rede 2	2001 - 2010	1084	1537	2.8358
Rede 3	2011 - 2018	1084	3831	7.0683

Tabela 1. Principais Características das Redes

Elkan [28], este problema interfere nos resultados devido a duas razões: (i) com menos exemplos de uma determinada classe, é mais difícil inferir padrões confiáveis; (ii) os modelos treinados são enviesados em direção a classe predominante. Diversos autores [5, 7, 17, 29] propõe técnicas e métodos para resolução deste desafio. 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 utilizada no trabalho aqui apresentado. Inicialmente 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 é o mesmo. Com os dados balanceados, o algoritmo para predição de ligações foi executado.

4 Resultados

Ao longo do processo descrito na seção anterior, o conjunto de dados sofreu algumas alterações. Os 1.094 pesquisadores podem possuir um total de 597.871 arestas. Destas, apenas 3.831 representavam arestas positivas na Rede 3, portanto, através do balanceamento das amostras, um conjunto randômico de outras 3.831 arestas ausentes foi escolhido. Sendo assim, o conjunto de dados utilizado para entrada no algoritmo de predição de dados é composto por 7.662 registros. Dessa forma, foram selecionados 5.746 ligações (escolhidas aleatoriamente) para treino, representando 25% do conjunto total, e outros 1.916 ligações para teste.

Diversos algoritmos podem ser utilizados para resolução de problemas de classificação, dentre estes, alguns foram selecionados para execução do trabalho, são eles: Regressão Logística, K-Vizinhos Mais Próximos, Baías Ingênuas e Florestas Aleatórias. Cada uma dessas técnicas apresenta uma particularidade diferente, e consequentemente, diferentes resultados. Portanto, seus resultados serão apresentados na Tabela 2, utilizando as métricas precisão, revogação, F1 e área sob a curva (AUC). Normalmente, em algoritmos utilizados para predição de ligações, a área sob a curva é utilizada pela maioria dos autores, portanto, a utilizaremos como base.

Cada uma das métricas utilizadas para validação dos resultados possui características próprias. A precisão tem como objetivo responder a seguinte pergunta: de todos os valores preditos positivos, quantos realmente estão corretos, uma alta precisão está relacionada a poucos falsos positivos. Já considerando todos os valores positivos, a revogação tem como objetivo saber quantos destes foram realmente preditos. A métrica F1 leva em consideração a precisão e a revogação, 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 ao longo de todo o processo de aprendizagem.

Analisando a Tabela 2, é possível perceber que os algoritmos escolhidos obtiveram bons resultados. Ao observar a área sob a curva, percebemos que todos obtiveram um resultado acima do que um mero acaso. Essa situação é melhor explicada na Figura 3, 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. Dessa forma, fica claro que o algoritmo conseguiu utilizar o conjunto de dados apresentado para realizar predições corretas a respeito de futuras ligações.

Algoritmo	**Precisão**	**Revogação**	**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

Tabela 2. Métricas geradas a partir das predições

Dentre os algoritmos utilizados, o que apresentou um 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 diferente entre os resultados obtidos, deixando claro que, para o problema em questão, ainda não podemos estabelecer qual técnica deveria ser utilizada como padrão.

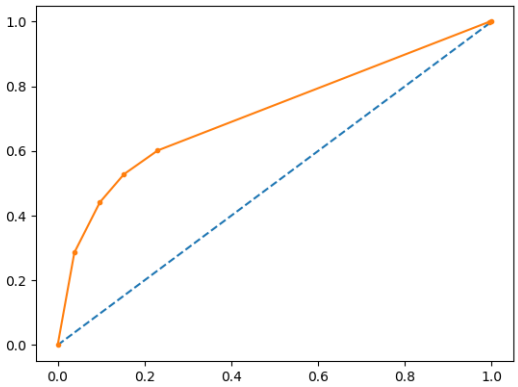


Figura 3. Área Sob a Curva (AUC) para o algoritmo K-Vizinhos Mais Próximos

5 Conclusão e trabalhos Futuros

Com objetivo de realizar a predição de futuras ligações em dados de colaboração científica, um conjunto de currículos de pesquisadores extraídos da Plataforma Lattes, mantida pelo CNPq, foi utilizado. Foram selecionados 1.091 currículos de pesquisadores com título de doutor, pertencentes à área de Ciências Sociais e Aplicas, e à sub-área Ciência da Informação. Três períodos de colaborações foram estabelecidos, sendo eles: 1960 à 2000, 2001 à 2010, e por fim, 2011 à 2018. As redes criadas para os períodos estabelecidos foram utilizadas para treino e teste dos modelos de classificação.

Para criação do conjunto de dados a ser inserido no algoritmo de aprendizagem de máquina, foram selecionados os atributos: Identificação do primeiro pesquisador, identificação do segundo pesquisador, CN (*Common Neighbors*), Coeficiente de Jaccard, Adamic/Adar, *Resource Allocation*,*Preferential Attachment*, Menor Caminho, peso, e por fim, a presença, ou não, de uma aresta. A partir disso, quatro algoritmos foram utilizados para realizar as predições. Os algoritmos apresentaram resultados similares uns aos outros, porém o que possui um melhor desempenho foi o K-Vizinhos Mais Próximos, apresentando precisão de 0.71, revogação igual a 0.68, F1 de 0.68 e área sob a curva de 0.71.

Os resultados aqui apresentados demonstram que é possível realizar a predição de ligações utilizando informações da própria rede estudada. O objetivo proposto foi então alcançado, uma vez que, a partir da utilização destes dados é possível saber, por exemplo, se dois pesquisadores da área citada acima, irão colaborar em um futuro instante de tempo. O desempenho das métricas de avaliação ficou em torno de 70% representando um bom resultado, porém valores maiores podem ser alcançados a partir da utilização de mais atributos.

Como 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. A partir disso, espera-se que os classificadores apresentem um desempenho ainda melhor.

Agradecimentos

Este trabalho é financiado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

References

[1] Newman, M. E., 2001. The structure of scientific collaboration networks. *Proceedings of the national academy of sciences*, vol. 98, n. 2, pp. 404–409.

[2] Newman, M. E. & Park, J., 2003. Why social networks are different from other types of networks. *Physical Review E*, vol. 68, n. 3, pp. 036122.

[3] Barabási, A.-L. & Albert, R., 1999. Emergence of scaling in random networks. *science*, vol. 286, n. 5439, pp. 509–512.

[4] Newman, M. E., 2003. Mixing patterns in networks. *Physical Review E*, vol. 67, n. 2, pp. 026126.

[5] Al Hasan, M. & Zaki, M. J., 2011. A survey of link prediction in social networks. In *Social network data analytics*, pp. 243–275. Springer.

[6] Liben-Nowell, D. & Kleinberg, J., 2003. The link-prediction problem for social networks. In *Conference on Information and Knowledge Management (CIKM'03)*, pp. 556–559.

[7] Acar, E., Dunlavy, D. M., & Kolda, T. G., 2009. Link prediction on evolving data using matrix and tensor factorizations. In *Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on*, pp. 262–269. IEEE.

[8] Zhou, T., Lü, L., & Zhang, Y.-C., 2009. Predicting missing links via local information. *The European Physical Journal B*, vol. 71, n. 4, pp. 623–630.

[9] Liu, Z., Zhang, Q.-M., Lü, L., & Zhou, T., 2011. Link prediction in complex networks: A local naïve bayes model. *EPL (Europhysics Letters)*, vol. 96, n. 4, pp. 48007.

[10] Maruyama, W. T. & Digiampietri, L. A., 2019. Co-authorship prediction in academic social network. In *Anais do V Workshop Brasileiro de Análise de Redes Sociais e Mineração*, pp. 79–90. SBC.

[11] Lane, J., 2010. Let's make science metrics more scientific. *Nature*, vol. 464, n. 7288, pp. 488.

[12] Dias, T. M., Moita, G. F., Dias, P. M., Moreira, T., & Santos, L., 2013. Modelagem e caracterização de redes científicas: um estudo sobre a plataforma lattes. In *BRASNAM-II Brazilian Workshop on Social Network Analysis and Mining*, pp. 10–20.

[13] Krebs, V. E., 2002. Mapping networks of terrorist cells. *Connections*, vol. 24, n. 3, pp. 43–52.

[14] Adamic, L. A. & Adar, E., 2003. Friends and neighbors on the web. *Social networks*, vol. 25, n. 3, pp. 211–230.

[15] Newman, M. E., 2004. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the national academy of sciences*, vol. 101, n. suppl 1, pp. 5200–5205.

[16] Dias, T. M. R., Moita, G. F., Dias, P. M., & Moreira, T. H. J., 2014. Identificação e caracterização de redes científicas de dados currículares. *iSys-Revista Brasileira de Sistemas de Informação*, vol. 7, n. 3, pp. 5–18.

[17] Perez-Cervantes, E., Mena-Chalco, J. P., De Oliveira, M. C. F., & Cesar, R. M., 2013. Using link prediction to estimate the collaborative influence of researchers. In *eScience (eScience), 2013 IEEE 9th International Conference on*, pp. 293–300. IEEE.

[18] Perez Cervantes, E., 2015. *Análise de redes de colaboração científica: uma abordagem baseada em grafos relacionais com atributos*. PhD thesis, Universidade de São Paulo.

[19] Mena-Chalco, J. P. & Junior, R. M. C., 2009. Scriptlattes: an open-source knowledge extraction system from the lattes platform. *Journal of the Brazilian Computer Society*, vol. 15, n. 4, pp. 31–39.

[20] Dias, T., 2016. Um estudo da produção científica brasileira a partir de dados da plataforma lattes. 181p. *Programa de Pós-Graduação em Modelagem Matemática e Computacional, Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte (Doutorado)*.

[21] Dias, T. M. R. & Moita, G. F., 2015. A method for the identification of collaboration in large scientific databases. *Em Questão*, vol. 21, n. 2, pp. 140–161.

[22] Digiampietri, L., Maruyama, W. T., Santiago, C., & da Silva Lima, J. J., 2015. Um sistema de predição de relacionamentos em redes sociais. In *Brazilian Symposium on Information Systems*, volume 11.

[23] Chen, H., Li, X., & Huang, Z., 2005. Link prediction approach to collaborative filtering. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'05)*, pp. 141–142. IEEE.

[24] Lü, L. & Zhou, T., 2011. Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications*, vol. 390, n. 6, pp. 1150–1170.

[25] Newman, M., 2010. Networks: An introduction oxford univ.

[26] Liben-Nowell, D. & Kleinberg, J., 2007. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, vol. 58, n. 7, pp. 1019–1031.

[27] Potgieter, A., April, K. A., Cooke, R. J., & Osunmakinde, I. O., 2009. Temporality in link prediction: Understanding social complexity. *Emergence: Complexity & Organization (E: CO)*, vol. 11, n. 1, pp. 69–83.

[28] Menon, A. K. & Elkan, C., 2011. Link prediction via matrix factorization. In *Joint european conference on machine learning and knowledge discovery in databases*, pp. 437–452. Springer.

[29] Gong, N. Z., Talwalkar, A., Mackey, L., Huang, L., Shin, E. C. R., Stefanov, E., Song, D., et al., 2011. Jointly predicting links and inferring attributes using a social-attribute network (san). *arXiv preprint arXiv:1112.3265*.