

Sugerindo colaborações na base de doutores da Plataforma Lattes usando Florestas Aleatórias

Gabriel Dahia, Gabriel L. P. e Souza, Pedro Vidal

¹Departamento de Ciência da Computação – Universidade Federal da Bahia (UFBA)

{gdahia, gabriellecomt, pvidal}@dcc.ufba.br

1. Introdução

A Plataforma Lattes, mantida pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), corresponde à integração de bases de dados de Currículos, de Grupos de Pesquisa e de Instituições. Ela pauta atividades de planejamento e gestão, e a formulação de políticas públicas dos órgãos governamentais brasileiros [CNPq 2010]. Em especial, o Currículo Lattes foi criado pelo CNPq para centralizar, padronizar e disponibilizar, através da Plataforma Lattes, informações pessoais, profissionais e acadêmicas, fornecidas pelos próprios autores dos currículos, da comunidade científica brasileira [Alves et al. 2011].

Apesar da riqueza e potencial dessa base de dados, o CNPq impõe restrições ao acesso dessas informações, limitando o seu estudo. Com intuito de disseminar o conteúdo da Plataforma Lattes, foi publicado o conjunto de dados *LattesDoctoralDataset*. Esta base contém dados a respeito do número e do tipo de publicações feitas, das colaborações entre pesquisadores, da atuação profissional e da formação acadêmica dos 265.187 doutores que possuem currículo publicado na plataforma [Dias et al. 2017].

Ao analisar as colaborações entre doutores nesse banco de dados, é possível perceber que seu número é de aproximadamente 0,02% do total possível. Com o intuito de fomentar colaborações entre pesquisadores e, por consequência, o avanço da ciência no Brasil, esse trabalho propõe, analisando o *LattesDoctoralDataset*, uma metodologia baseada em mineração de dados para sugerir novas colaborações com base nas informações fornecidas e nas colaborações preexistentes.

Usando Florestas Aleatórias [Ho 1995], o método proposto atinge acurácia de 86,70% \pm 0,18, com 99% de confiança, quando a tarefa é: dado um par de currículos, determinar se há pelo menos uma colaboração entre esses pesquisadores. Quando a tarefa é estimar o número de colaborações entre um par de pesquisadores, dados os seus currículos, o método consegue acurácia de 78,34% \pm 0,16, com 99% de confiança.

2. Metodologia

Para sugerir colaborações entre os doutores da base, primeiramente foi feito um pré-processamento na base de dados. Este envolveu remover currículos com informações inconsistentes ou faltantes, agrupar as localizações e dados sobre a formação dos doutores, campos de preenchimento livre na Plataforma Lattes, utilizando técnicas de mineração de texto, e a remoção de atributos irrelevantes para a tarefa proposta.

Com a base pré-processada, o modelo empregado foi de Florestas Aleatórias, um conjunto de Árvores de Decisão treinado com diferentes partes da base de treino.

caso de formacao, ou 3.000 grupos, no caso de lugares. Este número foi calculado para comportar todas instituições de ensino superior no Brasil - 2.364, de acordo com [Ministério da Educação 2016] - e outras.

Depois desse agrupamento, calculou-se a produtividade de cada doutor como o total de colaborações relatadas; esse campo foi denominado “Colaboracoes”. Utilizando esse atributo como preditor da tarefa que se deseja resolver, calculou-se, entre todo outro atributo X da base de dados e $Y = \text{Colaboracoes}$ a informação mútua I :

$$I(X, Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right), \quad (1)$$

onde $p(x, y)$ é a probabilidade conjunta de X e Y , e $p(x)$ e $p(y)$ são, respectivamente, as probabilidades marginais de X e Y . $p(x)$, $p(y)$ e $p(x, y)$ são estimados, respectivamente, como as frequências empíricas de x em X , y em Y , e (x, y) em (X, Y) .

Informalmente, $I(X, Y)$ quantifica quão independente as variáveis X e Y são; isso permite quantificar dependências lineares, como correlação é capaz de fazer, e não-lineares entre as distribuições estudadas. A figura 1 mostra os resultados.

Nessa etapa, todos os atributos independentes com “Colaboracoes” foram descartados do conjunto de dados. O resultado foi uma versão da base de dados em que, em uma determinada execução, restaram 8.180 currículos restritos aos atributos não independentes de “Colaboracoes”. Nesta versão da base, existem 58.689 colaborações entre pesquisadores, em contraste com as 6.902.042 colaborações da base original.

2.2. Modelo de mineração de dados

O método de Florestas Aleatórias pode ser descrito da seguinte maneira: são construídos N conjuntos de dados de tamanho igual ao da base original, através do sorteio com repetição dos exemplos do conjunto de dados inicial. Em cada uma dessas novas bases, é treinada uma Árvore de Decisão. O conjunto de todas essas árvores é o modelo de mineração de dados.

Para fazer inferência com esse modelo, são computadas as classes previstas por cada uma das árvores. A predição final é obtida através do cálculo ou da média ou da moda das previsões individuais.

Esse modelo pode equivalentemente ser explicado como uso de *bagging* para Árvores de Decisão. Suas vantagens: são regularizar o modelo, diminuindo sobreajuste e aumentando a acurácia [Ho 1995].

3. Experimentos

Como há estocasticidade no método a partir da etapa de pré-processamento, rodou-se o experimento 10 vezes com sementes aleatórias diferentes para cálculo de intervalos de confiança.

Para cada execução individual do experimento, a base pré-processada foi dividida em um conjunto de treino, compreendendo 60% do total de dados, e um conjunto de teste, com o restante. A divisão é feita de maneira estratificada, considerando apenas se há ou não colaboração entre os donos dos currículos.

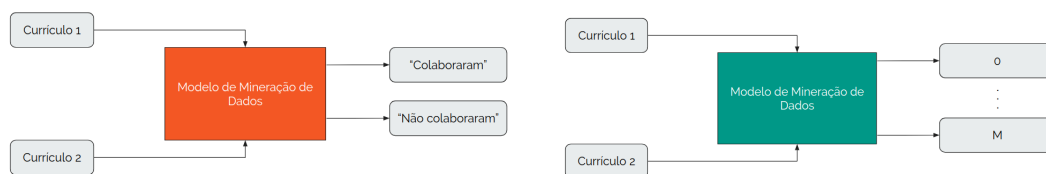


Figura 2. Modalidades de mineração de dados: à esquerda, a modalidade *binary*, em que é necessário apenas determinar se há ou não colaboração entre os donos dos currículos. À direita, a modalidade *standard*, onde o par de currículos é classificado de acordo com o número de colaborações previsto.

O restante do experimento consiste em treinar um modelo de Floresta Aleatória no conjunto de treino e validá-lo no conjunto de teste em duas modalidades distintas, que denominou-se *binary* e *standard*. A figura 2 ilustra ambas modalidades.

Na modalidade *binary*, a tarefa consiste em determinar, dado um par de currículos, se há pelo menos uma colaboração entre eles. Aqui, a Floresta Aleatória é treinada com pares de currículos como atributos e a informação de se há ou não colaboração entre eles como saída esperada. Para fazer com que o modelo treinado seja robusto à ordem em que os currículos são colocados no par, duplicou-se o conjunto de treino, colocando currículos nos dois sentidos.

Já na modalidade *standard*, a tarefa consiste em estimar, dado um par de currículos, o número de colaborações entre eles. Neste caso, o modelo é treinado com os pares de currículos e sua saída é uma classe de $\{0, 1, \dots, M\}$, onde M é o maior número de colaborações presente no conjunto de testes.

A validação do método usa a métrica acurácia, calculada como

$$acurácia = \frac{\text{Previsões corretas}}{\text{Total de previsões}}. \quad (2)$$

Assim como é feito para o conjunto de treino, pares de currículos são também colocados nos dois sentidos no conjunto de teste. Isso permite avaliar a robustez do método proposto à ordem em que os currículos são entregues ao modelo de mineração de dados sem artificialmente melhorar o resultado no conjunto de testes.

3.1. Resultados

A partir das dez repetições do experimento, obteve-se $86,70\% \pm 0,18$ de acurácia, com 99% de confiança, para a modalidade *binary*, e $78,34\% \pm 0,16$, com 99% de confiança, para a modalidade *standard*.

Quando se compara estes resultados com a performance esperada de um classificador aleatório para essa mesma tarefa - dado que esta tarefa não foi tentada anteriormente e, portanto, não há estado-da-arte -, respectivamente 50% para a modalidade *binary*, e 1,08% (em média, existem 92 valores de colaborações possíveis) para a modalidade *standard*.

Também é importante notar que a discrepância nos resultados entre as modalidades pode ser explicada pelo fato de a tarefa proposta para *standard* ser consideravelmente mais difícil do que a de *binary*: é possível que um dado número de colaborações

não apareça no treino e todas as instâncias dessa classe seriam mal-classificadas. Além do mais, a avaliação proposta não considera a proximidade do número de colaborações previsto do real; apenas previsões corretas são consideradas.

Apesar do grande número de nós em cada árvore (aproximadamente 20 mil), é possível realizar inferência no modelo, uma vez carregado em memória principal, em tempo real. Isso ocorre porque a altura é, no máximo, o número de atributos fornecidos ao modelo, que não passa de 100. Além disso, os arquivos dos modelos têm tamanho razoável: aqueles para a modalidade `binary` têm em torno de 15MB, e os `standard` têm em torno de 250MB.

4. Conclusão

Esse trabalho apresentou um modelo de mineração de dados baseado em Florestas Aleatórias capaz de prever colaborações entre doutores da base *LattesDoctoralDataset*. Além de estabelecer resultados iniciais para essa base, a acurácia obtida mostra que esse sistema é preciso o suficiente para ser utilizado na prática, sem problemas com tempo de execução ou requisito de armazenamento.

Foram identificadas, contudo, limitações da metodologia proposta que podem ser melhoradas em versões futuras. O tratamento de dados faltantes ou fora de conformação pode ser melhorado já que atualmente consiste em eliminar instâncias com erros detectáveis e mineração de texto para os campos de preenchimento livre.

A abordagem de mineração de textos empregada, inclusive, não generaliza caso haja termos não presentes no conjunto de dados analisado - também imagina-se que um método de mineração de textos mais robusto que LSA possa melhorar os resultados.

Melhores resultados também devem poder ser obtidos integrando o processo de colheita de uma nova versão da base de dados com as demandas do modelo de aprendizado de máquina. Como mencionado anteriormente, maior padronização das informações e a informação temporal das colaborações pode melhorar a performance da metodologia.

Referências

- Alves, A. D., Yanasse, H. H., and Soma, N. Y. (2011). Sucupira: a system for information extraction of the lattes platform to identify academic social networks. In *CISTI*, pages 1–6. IEEE.
- CNPq (2010). Sobre a Plataforma Lattes. <http://memoria.cnpq.br/web/portal-lattes/sobre-a-plataforma>.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *JASIST*, 41(6):391.
- Dias, T. M. R., Laender, A. H. F., and Moita, G. F. (2017). LattesDoctoralDataset: Uma coleção de dados estratificados sobre o conjunto de doutores cadastrados na Plataforma Lattes. In *SBBD*, pages 245–255.
- Ho, T. K. (1995). Random decision forests. In *International conference on document analysis and recognition*, volume 1, pages 278–282. IEEE.
- Ministério da Educação (2016). Altos índices de desistência na graduação revelam fragilidade do ensino médio, avalia ministro. <http://portal.mec.gov.br/ultimas-noticias/212/40111>.