Análise da relação entre obtenção de bolsas de produtividade do CNPq e medidas bibliométricas e de análise de redes sociais

Felipe P. C. Fonseca[1], **Luciano A. Digiampietri**[1]

[1]Escola de Artes, Ciências e Humanidades – Universidade de São Paulo (USP)
São Paulo – SP – Brazil

felipe.fonseca@usp.br, digiampietri@usp.br

Abstract. *There are different ways of encouraging scientific work. One of them is the research productivity grants given by CNPq. However, with the scarcity of resources, analyze the performance of researchers and select those who will receive the grants are becoming even more complex and challenging activities. Thus, this paper aims to build a classifier by using the Lattes Platform as a data source that can identify the grant level of a given researcher and help researchers to contextualize themselves in relation to other scholars, considering social network and bibliometric analyses.*

Resumo. *Existem diferentes formas de incentivo ao trabalho científico. Uma delas é a concessão de bolsas de produtividade em pesquisa por parte do CNPq. Porém, com a escassez de recursos, a análise do desempenho dos pesquisadores e seleção daqueles que receberão as bolsas são atividades cada vez mais complexas e desafiadoras. Desta forma, este artigo visa a construir um classificador, usando currículos da Plataforma Lattes como fonte de dados, que consiga identificar qual o nível da bolsa de um dado pesquisador e que o auxilie a se contextualizar em relação aos bolsistas considerando medidas bibliométricas e métricas oriundas da análise de redes sociais.*

1. Introdução e Motivação

Atualmente, o governo brasileiro incentiva a área acadêmica e a pesquisa por meio de instituições de fomento, tais como a CAPES e o CNPq. Dentre as diversas formas de se promover esse incentivo existe a modalidade de bolsas e auxílios, que tem como público alvo tanto alunos do ensino médio, alunos de graduação e de pós quanto doutores-pesquisadores. O foco desse artigo será nos doutores-pesquisadores.

A pequena disponibilidade de bolsas combinada com um cenário econômico desfavorável e certa subjetividade inerente ao processo torna a análise da produtividade acadêmica dos pesquisadores e a seleção dos que serão contemplados com as bolsas de produtividade em pesquisa atividades cada vez mais importantes e desafiadoras. Portanto, conseguir identificar pesquisadores com potencial para receber essas bolsas é uma tarefa relevante, pois pode ajudar o governo a distribuir de forma mais clara as bolsas fazendo com que o planejamento desse programa seja mais eficaz. Adicionalmente, não existem muitas informações para auxiliar um pesquisador que está pensando em pleitear uma dessas bolsas a identificar o contexto em que se encontra em relação aos demais pesquisadores e, em particular, àqueles possuidores de bolsa de produtividade em pesquisa.

Contudo, são poucas as pesquisas na área de identificação de potenciais bolsistas. Assim sendo, este projeto visa a construir um classificador usando os dados dos doutores na área de Ciências da Computação disponíveis a partir da plataforma Lattes[1] para identificar o nível da bolsa de um dado pesquisador, bem como permitir que um pesquisador consiga ter uma visão contextualizada de algumas de suas métricas em relação às métricas dos bolsistas.

2. Trabalhos Relacionados

Esta seção descreve trabalhos correlatos ao trabalho atual que focam tanto na avaliação de instituições de ensino e pesquisa como na avaliação de pesquisadores.

Uma análise da produtividade de uma instituição como um todo pode ser conferida em [Tess et al. 2009]. Esse artigo faz a avaliação da produção do Instituto do Coração, uma das 13 instituições da escola de medicina da Universidade de São Paulo. Destaca-se a abordagem usada, que se baseia em indicadores bibliométricos para fazer uma avaliação de toda a produção científica da instituição no período de janeiro de 2000 até dezembro de 2003.

Lima et al. (2015) fizeram a avaliação de bolsas produtividade e seu impacto sobre o desempenho dos bolsistas. Nesse trabalho é apresentada uma explicação detalhada de como um pesquisador aplica para uma bolsa e como ele é avaliado. Basicamente existem seis principais classes de bolsas produtividade em pesquisa fornecidas pelo CNPq: Sênior, 1A, 1B, 1C, 1D e 2.

Com o intuito de classificar um pesquisador candidato, cada comitê de avaliação analisa o projeto de pesquisa e o currículo do proponente. Tal currículo possui: suas publicações (artigos em revistas, artigos em conferências, livros, etc.); supervisão de estudantes (incluindo mestrado e doutorado); contribuições para a ciência, tecnologia e inovação (incluindo patentes); coordenação e participação de projetos de pesquisa; inserção internacional; participação como avaliador de conferências; e outras atividades relacionadas à ciência e à academia [Lima et al. 2015]. Contudo, os autores informam como índices de produtividade ou qualidade (por exemplo, o JCR e o Qualis usados no presente artigo) devem ser avaliados, respeitando as diversas áreas de atuação do pesquisador, a fim de ajudar a produzir uma análise da produtividade de uma forma mais imparcial e contextualizada.

Uma análise dos pesquisadores que possuem bolsas produtividade utilizando-se de uma rede social de coautorias para o cálculo das métricas já foi feita anteriormente [Andrade 2015]. Andrade e Régo mostram que métricas estruturais de redes sociais (de centralidade) têm uma correlação positiva com o nível do desempenho dos bolsistas. Conforme apresentado pelos autores "*as métricas de centralidade de grau, centralidade de proximidade, centralidade de intermediação e centralidade de autovetor são importantes atributos estruturais na rede social e estão relacionadas com a eficiência, liderança e satisfação*" [Andrade 2015].

Digiampietri et al. (2014) apresentam uma análise da participação dos orientandos na produção dos seus orientadores por meio de uma rede de coautoria (da mesma forma que o artigo citado anteriormente). O mesmo autor e seus colaboradores analisam

[1] http://lattes.cnpq.br/

redes sociais de pesquisadores por meio de subredes regionais de forma a se identificar as características de relacionamentos entre os doutores [Digiampietri et al. 2014a].

3. Metodologia

A metodologia usada se baseia na proposta por Digiampietri et al. (2013). No presente trabalho, algumas tarefas foram executadas (na forma de fases) a fim de se obter os dados necessários para a classificação dos pesquisadores. A primeira tarefa foi a obtenção de dados propriamente dita e a transformação dos mesmos (por meio de um *parser*) para torná-los mais fáceis de se manejar. Nessa fase as medidas bibliométricas de cada pesquisador foram obtidas.

Após isso, uma rede social foi construída (representada como um grafo) utilizando-se as relações de coautorias como arestas. Tal rede foi utilizada para a obtenção das medidas de centralidade. Por meio dos dados obtidos nas fases anteriores alguns testes, usando vários classificadores disponíveis no Weka, foram feitos para se avaliar o desempenho da classificação considerando o conjunto de medidas extraídas ou calculadas.

3.1. Obtenção e tratamento de dados

Foi escolhido como amostra o conjunto formado por todos os doutores que atuam em Ciência da Computação e possuem currículo Lattes[2]. Este conjunto é composto de 7.469 doutores. A fase de obtenção de dados é aquela responsável por transformar os dados brutos em formato XML dos Currículos Lattes dos doutores da amostra em um arquivo de texto tabulado, que posteriormente foi utilizado para montar o conjunto de dados. Note que nesse primeiro momento, as informações obtidas são referentes a cada pesquisador, sendo elas de carácter pessoal (nome, número de identificação do Lattes, entre outros) ou de carácter bibliométrico (número de publicações, número de orientações, etc.). Além disso, dados sobre as produções científicas foram obtidos (revistas e congressos em que cada artigo foi publicado, e os coautores) para serem usadas tanto na obtenção de índices de impacto referentes aos veículos de publicação quanto na construção da rede social de coautoria.

Foram analisados os dados de produção e a rede de coautoria considerando as publicações e orientações no período de 2010 e 2014. A relação de bolsistas produtividade foi extraída do site do CNPq, considerando os pesquisadores que possuíam bolsa vigente na área de Ciência da Computação em 2014.

3.2. Construção da rede social usando coautorias como arestas

Como citado no seção 2, a abordagem de se construir uma rede social usando coautorias como as arestas não é nova [Digiampietri et al. 2014a, Andrade 2015]. Nesta abordagem, a rede social é representada por um grafo não direcionado em que cada nó representa um pesquisador e cada aresta representa uma coautoria entre os pesquisadores (coautoria em publicações científicas).

Os algoritmos para resolução de entidades e construção da rede de coautoria utilizados neste trabalho foram aqueles disponibilizados por Digiampietri et al. (2014b). Já os algoritmos para o cálculo das métricas estudadas foram os disponibilizados por Digiampietri et al. (2016).

[2]http://lattes.cnpq.br/

3.3. Construção do conjunto e classificadores testados

Partindo-se dos dados coletados e/ou calculados nas fases anteriores, um conjunto de dados foi construído utilizando as medidas bibliométricas e de centralidade (considerando a rede social acadêmica [Digiampietri et al. 2012b]) de cada pesquisador [Wasserman and Galaskiewicz 1994, Lemieux and Ouimet 2008, Poblacion et al. 2009, Prell 2012]. Os atributos do conjunto podem ser consultados na tabela 1, junto das siglas que serão utilizadas no decorrer deste artigo.

Tabela 1. Atributos extraídos ou calculados.

Atributo	Descrição
Bolsista (BOL)	Adota valores entre 0 (o pesquisador não possui bolsa) ou 1 (ele possui).
Nível (NIV)	Adota valores entre 0 e 6. Além de representar se o pesquisador é bolsista ou não, denota o nível da bolsa do pesquisador (sendo 0 atribuído ao pesquisador sem bolsa; 1 ao pesquisador com bolsa nível 2; até 5 ao pesquisador 1A e 6 ao bolsista sênior.)
Dissertações de mestrado concluídas (DMC)	Representa o número de orientações em dissertações de mestrado concluídas.
Iniciações científicas concluídas (ICC)	Representa o número de orientações em iniciações científicas completas.
Monografias de conclusão de curso em aperfeiçoamento e especialização concluídas (MAC)	Representa o número de orientações em monografias de aperfeiçoamento e especialização concluídas.
Trabalhos de conclusão de curso concluídos (TCCC)	Número de orientações em TCCs concluídos.
Teses de doutorado completos (TDC)	Número de orientações em teses de doutorado concluídas.
Orientações de outra natureza concluídas (ONC)	Representa o número de orientações de qualquer natureza diferentes das citadas anteriormente, concluídas.
Supervisões de Pós-Doutorado concluídas (SPD)	Número de supervisões de pós-doutorado concluídas.
Iniciações científicas em andamento (ICA)	Número de orientações em iniciações científicas em andamento.
Trabalhos de conclusão de curso em andamento (TCCA)	Número de orientações em TCCs em andamento.
Monografias de conclusão de curso em aperfeiçoamento e especialização em andamento (MAA)	Representa o número de orientações em monografias de aperfeiçoamento e especialização em andamento.
Orientações de outra natureza em andamento (ONA)	Representa o número de orientações de qualquer natureza diferentes das citadas anteriormente em andamento.
Degree (DGR)	Medida de centralidade de grau da rede social construída na seção 3.2.
Betweenness (BTW)	Medida de centralidade de intermediação da rede social construída na seção 3.2
Closeness (ClS)	Medida de centralidade de proximidade da rede social construída na seção 3.2
Strength (STR)	Medida da força que o nó representado pelo pesquisador possuí na rede social.
PageRank (PRK)	Valor do PageRank que o nó do pesquisador possui na rede social.
Número de Eventos (NDE)	Número de artigos completos publicados em anais de eventos pelo pesquisador.
Número de Periódicos (NDP)	Número de artigos publicados em periódicos pelo pesquisador.
JCR Periódicos (JCR)	Valor do JCR para os periódicos nos quais pesquisador publicou.
Qualis Periódicos (QP)	Valor do Qualis para os periódicos nos quais o pesquisador publicou.
Qualis Conferências (QC)	Valor do Qualis para as conferências nas quais o pesquisador publicou.
Qualis Total (QT)	Valor do Qualis total para o pesquisador.

O próximo passo foi avaliar (por meio de um estudo exploratório) o desempenho de vários classificadores usando o conjunto de dados construído. Os classificadores avaliados foram os seguintes: *Naive Bayes*, Árvore de Decisão, *Random Tree*, Rede Bayesiana, *Adaboost* e *Decision Table*. Estes classificadores foram escolhidos por representarem classificadores de diferentes tipos/famílias.

4. Análise dos Resultados

Os resultados dos classificadores foram obtidos utilizando o arcabouço Weka[3] ou por meio de sua interface gráfica ou usando sua API num aplicativo desenvolvido em Java.

[3]http://www.cs.waikato.ac.nz/ml/weka/

É importante mencionar que dois teste foram feitos inicialmente sobre o conjunto. O primeiro avaliou o desempenho dos algoritmos usando como classe o nível da bolsa, e o outro avaliou se o pesquisador é bolsista ou não, sendo este o principal objetivo do trabalho.

Para o segundo caso, pelo fato do conjunto de dados ser desbalanceado (apenas 4,9% dos pesquisadores da amostra são bolsistas) foram realizados testes não balanceando e balanceando o conjunto de treinamento (utilizando a estratégia de *resampling*). Por esse motivo a API do Weka foi necessária, por não existir nenhuma implementação de *resampling* disponível no Weka da forma necessária ao presente trabalho. Tal aplicação computa a revocação (*recall*), Medida-F (*f-measure)* e precisão (*precision*) de cada classe para verificar se, mesmo com uma eventual perda de acurácia, o desempenho dos algoritmos melhora nesses outros indicadores. Todos os resultados apresentados nesta seção constituem a média dos resultados utilizando a estratégia de validação cruzada com 10 subconjuntos (*10-fold cross validation*).

Com o intuito de estruturar melhor os resultados, esta seção está organizada em quatro subseções: (i) uma que diz respeito aos resultados acerca do nível da bolsa; (ii) uma com resultados com respeito se o pesquisar é bolsista ou não mantendo o conjunto de dados original; (iii)uma que faz a mesma análise, mas utilizando-se da técnica de *resampling* para tentar obter melhores resultados; e (iv) uma avaliando o impacto que os atributos do conjunto de dados têm sobre os resultados dos classificadores, por meio do cálculo das correlações entre os mesmos e da análise de componentes principais (*Principal Component Analysis*).

4.1. Resultados - Nível da Bolsa

Os algoritmos testados nessa fase foram os seguintes: *Naive Bayes*, Árvore de Decisão, *Random Tree*, Rede Bayesiana e Adaboost. Em cada um deles as medidas de *acurácia*, *revocação*, *precisão* e *Medida-F* foram cálculos. Os resultados podem ser vistos nas subseções a seguir.

4.1.1. Árvore de Decisão

Os resultados para a árvore de decisão podem ser observados na tabela 2. Note que a acurácia geral deste classificador foi de $94,2964\%$, o que não é muito alta, visto que a classe dominante (não bolsistas) representa $95,06\%$ dos dados. Desta forma, classificar todos os doutores como não bolsistas teria um desempenho acima deste classificador (considerando apenas esta métrica).

4.1.2. *Naive Bayes*

Na tabela 3 são apresentados os resultados para o classificador Naive Bayes. A acurácia geral foi de $84,362\%$, sendo o pior resultado entre os classificadores. Porém, mesmo com a acurácia baixa, vemos a revocação para a classe Bolsa Tipo 2 consideravelmente alta, visto que não foi realizada nenhuma abordagem de balanceamento para esse teste.

Tabela 2. Resultados dos testes usando a técnica de validação cruzada com 10 subconjuntos (Árvore de Decisão).

classe	precisão	revocação	Medida-F
Não Bolsista	0,97	0,982	0,976
Bolsa Tipo 2	0,322	0,243	0,277
Bolsa Tipo 1D	0,098	0,08	0,088
Bolsa Tipo 1C	0,057	0,056	0,056
Bolsa Tipo 1B	0,182	0,083	0,114
Bolsa Tipo 1A	0,214	0,13	0,162
Bolsa Tipo SR	0	0	0

Tabela 3. Resultados dos testes usando a técnica de validação cruzada com 10 subconjuntos (*Naive Bayes*).

classe	precisão	revocação	Medida-F
Não Bolsista	0,987	0,894	0,939
Bolsa Tipo 2	0,212	0,545	0,305
Bolsa Tipo 1D	0,161	0,3	0,21
Bolsa Tipo 1C	0,114	0,278	0,161
Bolsa Tipo 1B	0,08	0,167	0,108
Bolsa Tipo 1A	0,026	0,043	0,033
Bolsa Tipo SR	0	0	0

4.1.3. *Random Forest*

A acurácia geral do algoritmo *Random Forest* foi de $95,1533\%$, o que é um pouco melhor do que os $95,06\%$ da classe dominante (não bolsistas). Uma limitação desse algoritmo e de praticamente todos avaliados nessa seção é que a revocação para as classes não dominantes é consideravelmente baixa (a melhor delas é de $16,6\%$).

Tabela 4. Resultados dos testes usando a técnica de validação cruzada em 10 subconjuntos (Random Forest).

classe	precisão	revocação	Medida-F
Não Bolsista	0,961	0,994	0,978
Bolsa Tipo 2	0,419	0,166	0,238
Bolsa Tipo 1D	0,278	0,1	0,147
Bolsa Tipo 1C	0	0	0
Bolsa Tipo 1B	0,286	0,083	0,129
Bolsa Tipo 1A	0,167	0,043	0,069
Bolsa Tipo SR	0	0	0

4.1.4. Rede Bayesiana

Assim como o Naive Bayes, a outra abordagem avaliada que utiliza o Teorema de Bayes como base para a classificação (Rede Bayesiana), não teve uma acurácia muito alta: $87,0933\%$. Porém, assim como com o Naive Bayes, observa-se a revocação das outras classes aumentar consideravelmente, sendo o melhor caso, para a classe Bolsa Tipo 2, com um valor de $62,6\%$.

Tabela 5. Resultados dos testes usando a técnica de validação cruzada em 10 subconjuntos (Rede Bayesiana).

classe	precisão	revocação	Medida-F
Não Bolsista	0,993	0,892	0,94
Bolsa Tipo 2	0,169	0,626	0,266
Bolsa Tipo 1D	0,16	0,32	0,213
Bolsa Tipo 1C	0,029	0,028	0,028
Bolsa Tipo 1B	0,042	0,083	0,056
Bolsa Tipo 1A	0,133	0,174	0,151
Bolsa Tipo SR	0	0	0

4.1.5. Adaboost

De todos os algoritmos avaliados, o que obteve o pior desempenho geral foi o Adaboost que classificou todos os pesquisadores como não bolsistas.

Tabela 6. Resultados dos testes usando a técnica de validação cruzada em 10-subconjuntos (Adaboost).

classe	precisão	revocação	Medida-F
Não Bolsista	0,951	1	0,975
Bolsa Tipo 2	0	0	0
Bolsa Tipo 1D	0	0	0
Bolsa Tipo 1C	0	0	0
Bolsa Tipo 1B	0	0	0
Bolsa Tipo 1A	0	0	0
Bolsa Tipo SR	0	0	0

4.2. Resultados - Bolsista ou não bolsista (sem *Resampling*)

Os resultados desta seção foram obtidos por meio de um aplicativo Java desenvolvido neste trabalho usando a API do Weka para acesso aos classificadores. Para os experimentos apresentados nesta seção não foi utilizada nenhuma estratégia de balanceamento do conjunto de treinamento, assim não são esperados bons resultados de revocação (*recall*) da classe bolsista, visto que as classes são desbalanceadas. Os resultados podem ser vistos na tabela 7.

O classificador que obteve a maior acurácia foi o *Rotation Forest* (96,18%), obtendo também a maior precisão para a classe positiva (65,55%). Em termos de revocação, destaca-se a Rede Bayesiana (87,80%). Por fim, o algoritmo que obteve a maior medida-F da classe positiva, medida que tenta ponderar a precisão e a revocação, foi o Adaboost, com 0,5544 (valor muito próximo ao obtido pelo classificador *Rotation Forest*).

4.3. Resultados - Bolsista ou não bolsista (com *resampling*)

Nesta seção são apresentados os resultados da classificação usando *resampling* como estratégia de balanceamento do conjunto de treinamento. Os valores de *recall*, *f-measure*, *precision* (por classe) e acurácia global foram calculados, usando como base os falsos positivos e negativos e os verdadeiros positivos e negativos. Pode-se conferir os resultados na tabela 8.

Tabela 7. Resultados de todos os classificadores usando o fato de ser bolsista ou não como classe (sem *resampling*).

		Recall	Precision	F-Measure	Acurácia
Árvore de Decisão	F	0,9803	0,97261	0,97643	0,9550
	T	0,4688	0,5527	0,5073	
Random Forest	F	0,9869	0,9733	0,9801	0,9618
	T	0,4797	0,6555	0,5540	
Naive Bayes	F	0,9290	0,9858	0,9565	0,9198
	T	0,7425	0,3521	0,4778	
Rede Bayesiana	F	0,9015	0,9930	0,9451	0,9004
	T	0,8780	0,3167	0,4655	
Adaboost	F	0,9667	0,9804	0,9735	0,9501
	T	0,6287	0,4957	0,5544	
Decision Table	F	0,9801	0,9727	0,9764	0,9550
	T	0,4715	0,5523	0,5087	

Tabela 8. Resultados de todos os classificadores usando o fato de ser bolsista ou não como classe (com *resampling*).

		Recall	Precision	F-Measure	Acurácia
Árvore de Decisão	F	0,9446	0,9832	0,9636	0,9321
	T	0,6910	0,3935	0,5015	
Random Forest	F	0,9613	0,9873	0,9741	0,9514
	T	0,7615	0,5054	0,6076	
Naive Bayes	F	0,9117	0,9871	0,9479	0,9048
	T	0,7723	0,3125	0,4450	
Rede Bayesiana	F	0,8880	0,9943	0,9381	0,8887
	T	0,9024	0,2952	0,4449	
Adaboost	F	0,8939	0,9953	0,9419	0,8952
	T	0,9187	0,3104	0,4641	
Decision Table	F	0,8666	0,9935	0,9257	0,8679
	T	0,8915	0,2578	0,3999	

Conforme esperado, o balanceamento do conjunto de treinamento utilizando *resampling* acarretou num aumento na revocação da classe positiva (isto é, bolsistas), porém com uma diminuição de precisão da classe positiva e também da acurácia geral. Destaca-se a revocação da classe positiva de 91,87% obtida pelo classificador Adaboost. Este tipo de abordagem é útil quando se deseja aumentar a revocação (com eventual perda de precisão) para, por exemplo, maximizar as chances dos indivíduos da classe minoritária serem efetivamente classificados como tal.

4.4. Análise dos Atributos

Para realizar uma análise detalhada sobre os atributos e suas influências sobre o resultado dos classificadores, serão usadas as correlação entre os atributos e também será realizada a Análise de Componentes Principais (*Principal Component Analysis*).

A correlação entre todos os atributos usados na classificação pode ser vista na tabela 9. Observe que os dois primeiros atributos são, respectivamente, o fato de o pesquisador ser bolsista ou não (0 para caso não seja bolsista e 1 caso seja) e o nível da bolsa (seguindo a mesma representação da seção 4.1).

Como pode ser observado na primeira linha da tabela 9, as medidas de centrali-

Tabela 9. Correlação entre os atributos

	BOL	NIV	DMC	ICC	MAC	TCCC	TDC	ONC	SPD	ICA	TCCA	MAA	ONA	DGR	BTW	CLS	STR	PRK	DNE	NDP	JCR	QP	QC	QT
BOL	1,000	0,813	0,302	-0,031	0,011	0,095	0,404	-0,016	0,139	0,044	-0,014	-0,015	-0,005	0,418	0,381	0,196	0,418	0,401	0,333	0,213	0,100	0,532	0,398	0,526
NIV	0,813	1,000	0,244	-0,030	-0,013	0,065	0,400	-0,015	0,125	0,022	-0,019	-0,014	-0,004	0,378	0,342	0,160	0,378	0,365	0,290	0,226	0,110	0,552	0,351	0,507
DMC	0,302	0,244	1,000	0,054	0,187	0,302	0,551	0,062	0,166	0,148	0,033	0,010	0,013	0,431	0,376	0,285	0,431	0,474	0,647	0,377	0,087	0,331	0,422	0,438
ICC	-0,031	-0,030	0,054	1,000	0,157	0,055	-0,020	0,047	-0,010	0,025	0,076	0,177	-0,005	-0,016	-0,009	-0,014	-0,016	-0,011	0,072	0,043	-0,023	-0,042	-0,034	-0,043
MAC	0,011	-0,013	0,187	0,157	1,000	0,393	0,033	0,280	0,005	0,121	0,223	0,040	-0,010	0,079	0,070	0,076	0,079	0,097	0,206	0,098	-0,021	0,001	0,058	0,037
TCCC	0,095	0,065	0,302	0,055	0,393	1,000	0,143	0,371	0,075	0,356	0,100	0,022	-0,010	0,155	0,128	0,144	0,155	0,193	0,340	0,216	0,040	0,118	0,152	0,157
TDC	0,404	0,400	0,551	-0,020	0,033	0,143	1,000	0,013	0,343	0,070	-0,041	-0,021	-0,009	0,443	0,435	0,215	0,443	0,494	0,549	0,454	0,189	0,462	0,431	0,511
ONC	0,016	0,015	0,062	0,047	0,280	0,371	0,013	1,000	0,018	0,120	0,084	0,028	0,005	0,024	0,023	0,018	0,024	0,031	0,106	0,073	0,009	0,011	0,027	0,012
SPD	0,139	0,125	0,166	-0,010	0,005	0,075	0,343	0,018	1,000	0,043	-0,022	-0,009	-0,003	0,124	0,126	0,059	0,124	0,157	0,163	0,282	0,228	0,184	0,100	0,158
ICA	0,044	0,022	0,148	0,025	0,121	0,356	0,070	0,120	0,043	1,000	0,148	0,030	-0,009	0,074	0,056	0,092	0,074	0,095	0,180	0,118	0,043	0,064	0,082	0,085
TCCA	-0,014	-0,019	0,033	0,076	0,223	0,100	-0,041	0,084	-0,022	0,148	1,000	0,084	-0,006	0,035	0,019	0,062	0,035	0,042	0,071	0,013	-0,020	-0,016	0,009	-0,002
MAA	-0,015	-0,014	-0,010	0,177	0,040	0,022	-0,021	0,028	-0,009	0,030	0,084	1,000	-0,002	-0,016	-0,014	-0,006	-0,016	-0,014	0,006	0,005	-0,011	-0,021	-0,017	-0,022
ONA	-0,005	-0,004	-0,013	-0,005	-0,010	-0,010	-0,009	-0,005	-0,003	-0,009	-0,006	-0,002	1,000	-0,013	-0,007	-0,024	-0,013	-0,018	-0,014	-0,009	-0,003	-0,008	-0,009	-0,010
DGR	0,418	0,378	0,431	-0,016	0,079	0,155	0,443	0,024	0,124	0,074	0,035	-0,016	-0,013	1,000	0,805	0,546	1,000	0,910	0,575	0,268	0,083	0,498	0,573	0,619
BTW	0,381	0,342	0,376	-0,009	0,070	0,128	0,435	0,023	0,126	0,056	0,019	-0,014	-0,007	0,805	1,000	0,316	0,805	0,750	0,523	0,250	0,072	0,444	0,510	0,552
CLS	0,196	0,160	0,285	-0,014	0,076	0,144	0,215	0,018	0,059	0,092	0,062	-0,006	-0,024	0,546	0,316	1,000	0,546	0,644	0,369	0,155	0,029	0,255	0,322	0,336
STR	0,418	0,378	0,431	-0,016	0,079	0,155	0,443	0,024	0,124	0,074	0,035	-0,016	-0,013	1,000	0,805	0,546	1,000	0,910	0,575	0,268	0,083	0,498	0,573	0,619
PRK	0,401	0,365	0,474	-0,011	0,097	0,193	0,494	0,031	0,157	0,095	0,042	-0,014	-0,018	0,910	0,750	0,644	0,910	1,000	0,615	0,334	0,111	0,487	0,548	0,598
NDE	0,333	0,290	0,647	0,072	0,206	0,340	0,549	0,106	0,163	0,180	0,071	0,006	-0,014	0,575	0,523	0,369	0,575	0,615	1,000	0,381	0,057	0,414	0,638	0,617
NDP	0,213	0,226	0,377	0,043	0,098	0,216	0,454	0,073	0,282	0,118	0,013	0,005	-0,009	0,268	0,250	0,155	0,268	0,334	0,381	1,000	0,756	0,576	0,227	0,440
JCR	0,100	0,110	0,087	-0,023	-0,021	0,040	0,189	-0,009	0,228	0,043	-0,020	-0,011	-0,003	0,083	0,072	0,029	0,083	0,111	0,057	0,756	1,000	0,508	0,070	0,305
QP	0,532	0,552	0,331	-0,042	0,001	0,118	0,462	-0,011	0,184	0,064	-0,016	-0,021	-0,008	0,498	0,444	0,255	0,498	0,487	0,414	0,576	0,508	1,000	0,510	0,838
QC	0,398	0,351	0,422	-0,034	0,058	0,152	0,431	0,027	0,100	0,082	0,009	-0,017	-0,009	0,573	0,510	0,322	0,573	0,548	0,638	0,227	0,070	0,510	1,000	0,897
QT	0,526	0,507	0,438	0,043	0,037	0,157	0,511	0,012	0,158	0,085	0,002	0,022	0,010	0,619	0,552	0,336	0,619	0,598	0,617	0,440	0,305	0,838	0,897	1,000

dade da rede social (especialmente, grau, intermediação, *strength* e *PageRank*) junto do número de orientações concluídas de doutorado e as medidas Qualis são aquelas com maior correlação com os dois atributos usados na classificação. Em particular, a maior correlação foi identificada com o atributo Qualis Periódicos (QP). Assim sendo, estes atributos são aqueles que potencialmente têm grande influência sobre a classificação.

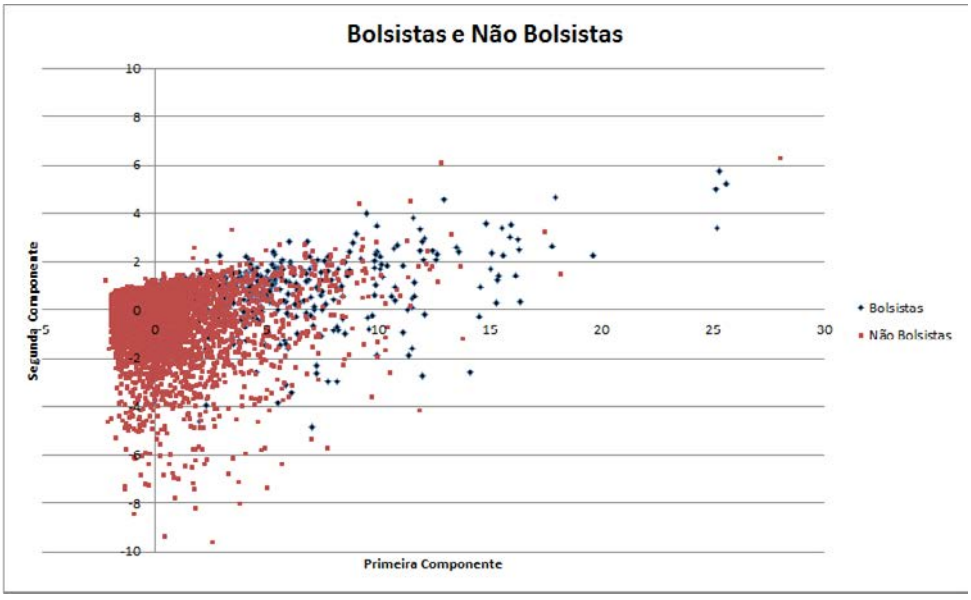

Figura 1. Plotagem das duas primeiras componentes resultantes do PCA

O próximo resultado corresponde ao gerado pela PCA (*Principal Component Analysis*). Usando os valores das duas primeiras componentes resultantes dessa abordagem como entradas para um gráfico e colorindo bolsistas e não bolsistas com uma cor diferente, tem-se as figuras 1, 2 e 3. O interessante de tais imagens é que pode-se obser-

var que os doutores não bolsistas concentram-se perto da origem do gráfico, enquanto que os bolsistas tendem a ficar mais espalhados e, em especial, com valores positivos para o eixo x do gráfico (que corresponde à componente principal resultantes da PCA).

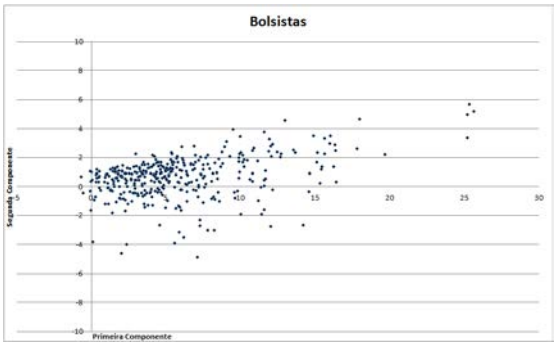

Figura 2. Plotagem das duas primeiras componentes resultantes do PCA - Bolsistas

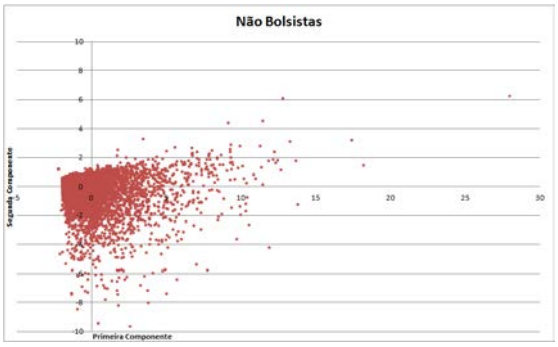

Figura 3. Plotagem das duas primeiras componentes resultantes do PCA - Não Bolsistas

Além disso, para auxiliar um dado pesquisador a conseguir se contextualizar mais facilmente em relação aos outros é gerado um gráfico posicionando as métricas desse pesquisador (aqui exemplificado por um pesquisador chamado de Pesquisador X) em relação aos demais bolsistas (note que o Pesquisador X pode ser qualquer pesquisador da área de Ciências da Computação). Essa análise é feita sobre os seus atributos e consegue mostrar o quão eficiente ou não este está em relação aos seus semelhantes, para isto, cada atributo (ou métrica) é normalizado para valores entre zero e um e a mediana e a média para cada atributo dos bolsistas produtividade são calculadas. Um exemplo desta contextualização é apresentado na figura 4.

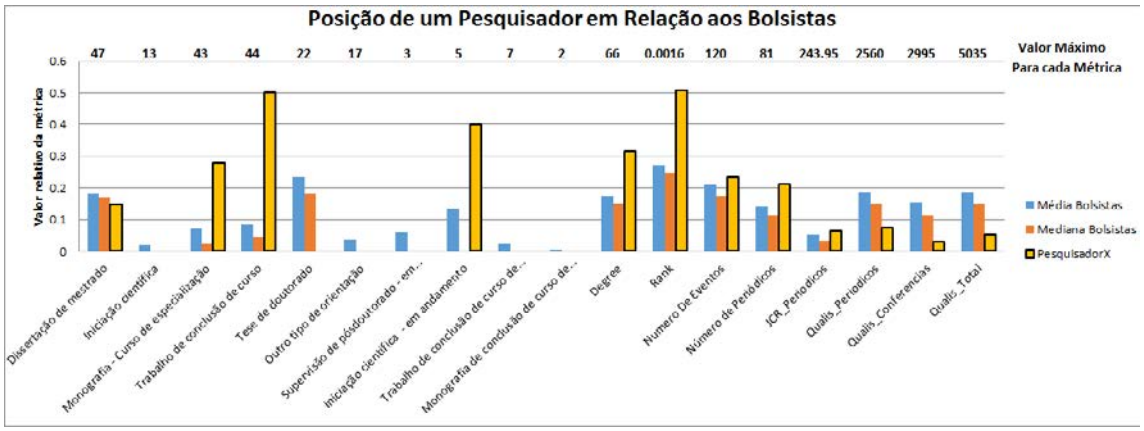

Figura 4. Contextualização das métricas de um pesquisador em relação às métricas dos bolsistas

5. Conclusões e Trabalhos Futuros

O processo de seleção de bolsistas de produtividade em pesquisa do CNPq é um processo complexo que envolve características objetivas e subjetivas que, atualmente, só é possível com o auxílio de milhares de pareceristas *ad-hoc*.

Como pôde ser observado ao se utilizar apenas algumas características objetivas (calculadas a partir de medidas bibliométricas e da análise de redes sociais), a maioria dos classificadores não obteve um desempenho melhor (em questão de acurácia) do que classificar todos os pesquisadores como não bolsistas, ou seja, nenhum deles conseguiu ultrapassar de forma consistente o valor de $95,06\%$ para a acurácia. Porém, é importante destacar que há outra medida importante para a classificação de bolsistas que é a revocação (*recall*).

Observa-se que na classificação entre bolsistas e não bolsistas as medidas de revocação da classe positiva (bolsistas) e precisão para esta mesma classe possuem resultados complementares. Enquanto o melhor resultado de precisão para esta classe foi de 65,5% (Rotation Forest) com revocação de pouco menos de 48%, tem-se no outro extremo os resultados da Rede Bayesiana com revocação da classe positiva superior a 87%, mas precisão pouco acima de 31%. Isto é, o algoritmo foi capaz de identificar mais de 87% dos bolsistas, porém do total de indivíduos que o algoritmo indicou como bolsistas, menos de 32% deles realmente eram. Por isso, entre os algoritmos testados, o classificador que utiliza a Rede Bayesiana foi aquele que apresentou a pior acurácia geral para o caso de identificação entre bolsista e não bolsista sem *resampling* (90,04%).

Uma estratégia comumente utilizada para tentar aumentar a revocação das classes minoritárias em conjunto de dados não balanceados é o balanceamento do conjunto de treinamento. Para isto, foram realizados os testes com *resampling*. Nestes, a revocação das classes minoritárias para todos os algoritmos subiu consideravelmente, enquanto o valor de acurácia diminuiu (fato que já era esperado). Todos os resultados obtiveram uma revocação de pelo menos 69% para a classe minoritária, mostrando que fazer o *resampling* funcionou de maneira satisfatória em relação a esta medida. A maior revocação foi obtida pelo algoritmo que utiliza Adaboost, chegando a um valor de $91,87\%$, com precisão para a classe positiva de 31% e acurácia geral de 89,52%

O intuito desse trabalho foi criar uma abordagem que conseguisse classificar pesquisadores da área de Ciências da Computação em bolsistas e não bolsistas além de facilitar a contextualização de um pesquisador em relação aos bolsistas e, conforme apresentado, considera-se que os objetivos foram atingidos de maneira satisfatória. Utilizando-se do Adaboost com *resampling*, por exemplo, foi obtida uma acurácia de $89,52\%$ e uma revocação para ambas as classes de aproximadamente 90%. Este tipo de classificação, mais do que verificar se efetivamente um doutor é ou não bolsista de produtividade, pode ser utilizada para indicar a um doutor que, de acordo com as métricas estudadas, ele possui melhores ou piores chances de obter uma bolsa de produtividade em pesquisa.

Como trabalhos futuros pretende-se explorar outras medidas bibliométricas e estruturais de redes sociais visando a obter resultados ainda melhores.

Agradecimentos

O trabalho apresentado neste artigo foi parcialmente financiado pela CAPES e pelo CNPq (processos 306046/2013-0 e 477246/2013-3).

Referências

Andrade, R. L.; Rego, L. C. (2015). A influência da rede de coautoria no nível das bolsas de produtividade da Área de Engenharia de Produção. In *IV Brazilian Workshop on*

Social Network Analysis and Mining (BraSNAM 2015).

Digiampietri, L., Linden, R., and Barbosa, L. (2016). Caracterizando departamentos e programas de computação utilizando análise de redes sociais e bibliometria. In *V Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2016)*.

Digiampietri, L., Mena-Chalco, J., Pérez-Alcázar, J. J., Tuesta, E. F., Delgado, K., and Mugnaini, R. (2012a). Minerando e caracterizando dados de currículos Lattes. In *III Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2014), Anais do XXXIV Congresso da Sociedade Brasileira de Computação (CSBC2014)*.

Digiampietri, L. A., Alves, C. M., Trucolo, C. C., and Oliveira, R. A. C. (2014a). Análise da rede dos doutores que atuam em computação no Brasil. In *III Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2014), Anais do XXXIV Congresso da Sociedade Brasileira de Computação (CSBC2014)*.

Digiampietri, L. A., Mena-Chalco, J., Silva, G. S., Oliveira, L., Malheiro, A., and Meira, D. (2012b). Dinâmica das relações de coautoria nos programas de pós-graduação em computação no Brasil. In *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2012) - Anais do XXXII Congresso da Sociedade Brasileira de Computação (CSBC 2012)*, page 12, Curitiba, PR, Brasil.

Digiampietri, L. A., Mena-Chalco, J. P., Vaz de Melo, P. O. S., Malheiro, A. P. R., Meira, D. N. O., Franco, L. F., and Oliveira, L. B. (2014b). BraX-Ray: An X-Ray of the Brazilian Computer Science Graduate Programs. *PLoS ONE*, 9(4):e94541.

Digiampietri, L. A., Mugnaini, R., and Alves, C. M. (2013). Análise da participação dos orientandos na produção dos orientadores: um estudo de caso em ciência da computação. In *II Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2013) - Anais do XXXIII Congresso da Sociedade Brasileira de Computação (CSBC 2013).*

Lemieux, V. and Ouimet, M. (2008). *Análise Estrutural das Redes Sociais*. Instituto Piaget.

Lima, H., Silva, T. H., Moro, M. M., Santos, R. L., Meira, Jr, W., and Laender, A. H. (2015). Assessing the profile of top Brazilian computer science researchers. *Scientometrics*, 103(3):879–896.

Poblacion, D., Mugnaini, R., and Ramos, L. (2009). *Redes sociais e colaborativas em informação científica*. Angellara Editoras, Sao Paulo, 1st edition.

Prell, C. (2012). *Social network analysis history, theory & methodology*. Los Angeles London SAGE.

Tess, B. H., Furuie, S. S., Castro, R. C. F., Barreto, M. d. C. C., and Nobre, M. R. C. (2009). Assessing the scientific research productivity of a Brazilian healthcare institution: a case study at the heart institute of São Paulo, Brazil. *Clinics*, 64:571 – 576.

Wasserman, S. and Galaskiewicz, J. (1994). *Advances in social network analysis research in the social and behavioral sciences*. SAGE.