Combinando Análise Bibliométrica e Análise de Redes Sociais para a Avaliação de Grupos Acadêmicos

Lucas Leal Caparelli[1], **Luciano Antonio Digiampietri**[1]

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

lucas.caparelli@usp.br, digiampietri@usp.br

Abstract. *The characterization and evaluation of groups of researchers are relevant and complex activities. This paper aims to characterize Brazilian graduate programs in Computer Science according to different bibliometric and social networks analysis metrics. In order to do this, the four-year evaluation of the programs carried out by CAPES was taken as the object of study, trying to identify which classification method is the most appropriate for this task. Using machine learning algorithms, it was possible to produce a classification model of these programs, reaching an accuracy of 86.15%.*

Resumo. *A caracterização e a avaliação de grupos de pesquisadores são atividades relevantes e complexas. Este artigo visa a caracterizar programas brasileiros de pós-graduação em Ciência da Computação de acordo com diferentes medidas bibliométricas e oriundas da análise de redes sociais. Para tal, tomou-se como objeto de estudo a avaliação quadrienal dos programas feita pela CAPES, buscando identificar qual método de classificação é o mais indicado para esta tarefa. Utilizando algoritmos de aprendizado de máquina foi possível produzir um modelo de classificação destes programas alcançando acurácia de 86,15%.*

1. Introdução

A avaliação de grupos acadêmicos é extremamente importante para tarefas como a concessão de financiamentos, análise da viabilidade de projetos de pesquisa, entre outras. Essa avaliação combina diversas informações diferentes, muitas das quais podem estar distribuídas em fontes de dados distintas, sendo muitas vezes subjetivas ou difíceis de quantificar.

Na pós-graduação brasileira, a CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) realiza periodicamente a avaliação de todos os programas nacionais de pós-graduação e os resultados dessa avaliação indicam se um programa está apto ou não a oferecer turmas de mestrado e/ou doutorado, bem como indicam a quantidade de recursos federais (para ajudar no financiamento do programa) e a quantidade de bolsas que serão reservados a cada programa.

Nas avaliações, cada uma das áreas de conhecimento detalha os critérios e pesos utilizados. Tipicamente, estes critérios são divididos em cinco eixos/quesitos: 1) *Proposta do Programa* que inclui coerência, planejamento e infraestrutura; 2) *Corpo Docente* incluindo o perfil, titulação, distribuição na execução das atividades e dedicação/regime de

trabalho; 3) *Corpo Discente* que analisa a quantidade de teses e dissertações defendidas, distribuição das orientações entre os docentes, qualidade da produção dos discentes e eficiência do programa na formação de mestres e doutores, 4) *Produção Intelectual* que analisa a quantidade e qualidade das publicações produzidas pelos discentes e docentes do programa, a distribuição das publicações entre os docentes, a participação de discentes nas publicações, e outros tipos de produções intelectuais (por exemplo, patentes); 5) *Inserção Social* que avalia a inserção e o impacto do programa, integração com outros programas e a visibilidade do programa.

Ao longo dos últimos anos, diversos trabalhos tentaram caracterizar de maneira automática grupos acadêmicos (como programas de pós-graduação, grupos de pesquisa ou departamentos), ou inferir o resultado de avaliações (de ranqueamentos nacionais ou internacionais, ou de programas de pós-graduação) [Digiampietri et al. 2014, Mena-Chalco et al. 2014, Digiampietri et al. 2016, Silva et al. 2017, Linden et al. 2017].

O objetivo do presente trabalho é combinar análise bibliométrica com análise de redes sociais a fim de se analisar a importância de diferentes atributos (ou métricas) na avaliação dos programas brasileiros de pós-graduação em Ciência da Computação, bem como avaliar a capacidade desses atributos na inferência dos conceitos atribuídos pela CAPES. Este trabalho se diferencia dos trabalhos correlatos por analisar dados da última avaliação da CAPES (quadriênio 2013 a 2016) e por combinar informações de três fontes diferentes: CAPES (Plataforma Sucupira), CNPq (Plataforma Lattes) e Google Acadêmico (*Google Scholar*).

2. Materiais e métodos

Neste trabalho foram analisados os 66 programas acadêmicos de pós-graduação em Ciência da Computação que foram avaliados pela CAPES na última avaliação quadrienal (período de 2013 a 2016) e que iniciaram suas atividades anteriormente ao período de avaliação.

Dos 66 programas avaliados, dois não possuíam conceito atribuído na avaliação anterior, 45 mantiveram o mesmo conceito do triênio anterior, 16 tiveram sua avaliação elevada em um ponto e três tiveram seu conceito reduzido em um ponto.

As informações básicas dos programas (nome, conceito CAPES do último quadriênio, conceito CAPES do triênio anterior e lista de docentes) foi obtida manualmente a partir da Plataforma Sucupira[1] no dia 15 de outubro de 2017. Com base na lista dos orientadores, foram identificados os currículos Lattes de cada um dos orientadores por meio de um processo automático [Digiampietri et al. 2012, Digiampietri et al. 2014] que foi posteriormente verificado manualmente para assegurar a correta identificação dos currículos. Ao todo foram identificados 1.608 currículos de orientadores.

Adicionalmente, para cada orientador foi buscado o respectivo perfil do Google Acadêmico[2] utilizando um procedimento automático para identificação de perfis [Digiampietri et al. 2014, Digiampietri and Ferreira 2018] nos dias 15 e 16 de novembro de 2017. Este processo identificou o perfil de 1.093 orientadores.

[1]Plataforma Sucupira: https://sucupira.capes.gov.br/sucupira/ , acessado em 29/01/2018

[2]Google Acadêmico: https://scholar.google.com.br/ , acessado em 29/01/2018

Dois tipos de informações foram extraídas de cada programa: informações bibliométricas e métricas da análise de redes sociais. A tabela 1 contém a lista e uma breve descrição dos 29 atributos bibliométricos utilizados. Os três primeiros atributos, gerais de cada programa, foram obtidos com base nos dados da Plataforma Sucupira. Os 14 atributos relacionados a orientações e publicações foram obtidos a partir de dados dos currículos da Plataforma Lattes, destacando-se que os atributos relacionados a "pontuações" das publicações combinaram dados das publicações com os valores associados a cada conceito Qualis definidos pelo Comitê de Área da Ciência da Computação. Por fim, os 12 últimos atributos desta tabela, relacionados às citações, foram extraídos dos perfis do Google Acadêmico.

A tabela 2 contém a lista e breve descrição dos 14 atributos oriundos da análise de redes sociais. No presente trabalho dois tipos de redes sociais baseadas nas relações de coautoria foram construídas (utilizando as informações das publicações extraídas dos currículos Lattes dos orientadores de cada programa). O primeiro tipo, composto por apenas uma rede, possui como nós os programas de pós-graduação e como arestas as relações de coautoria entre docentes de diferentes programas. Esta rede é utilizada para o cálculo de quatro métricas de centralidade (ou importância) de cada programa em relação às ligações com os demais. Adicionalmente, foi construída uma rede por programa, na qual cada nó corresponde a um orientador e cada aresta corresponde a relações de coautoria entre orientadores de um mesmo programa. A partir da rede de cada programa foram extraídas 10 métricas globais de rede.

Com base nos atributos extraídos ou calculados, o objetivo deste trabalho foi analisar a importância desses atributos em relação ao Conceito CAPES obtido pelos programas na última avaliação quadrienal. Dois tipos de análise foram realizadas com este propósito: análise da importância dos atributos em relação ao Conceito CAPES (baseada na correlação de valores e em algoritmos de seleção de atributos) e a capacidade de se inferir o Conceito CAPES com base nos demais atributos utilizando algoritmos de classificação.

Tanto para a seleção de atributos quanto para a classificação dos programas de pós-graduação em Ciência Computação das universidades brasileiras utilizou-se o arcabouço Weka, o qual foi criado pela Universidade de Waikato, na Nova Zelândia [Witten et al. 2016]. Este arcabouço permite acesso a diversos métodos de classificação, dos quais fez-se uso em testes com o conjunto de dados a fim de se identificar o modelo de maior acurácia.

Para seleção de atributos fez-se uso de dois seletores de atributos: *ChiSquaredAttributeEval* e *CFSSubsetEval*. O seletor de atributos *ChiSquaredAttributeEval* avalia a importância de um atributo a partir da estatística de inferência qui-quadrado que serve para avaliar quantitativamente a relação entre um resultado (no caso o conceito CAPES atribuído aos programas) e a distribuição dos valores de um dado atributo. O método para a identificação dos atributos mais importantes utilizado foi o *Ranker* que, neste caso, simplesmente ordena os atributos de acordo com o respectivo valor de qui-quadrado.

O seletor *CFSSubsetEval* avalia a qualidade de um subconjunto de atributos considerando sua habilidade de predição individual juntamente do grau de redundância entre os atributos deste subconjunto. Subconjuntos que possuem alta correlação com o atributo

Tabela 1. Atributos bibliométricos utilizados

Sigla utilizada	Descrição
Conceito CAPES	Conceito CAPES do programa para o quadriênio 2013-2016.
Conceito CAPES Anterior	Conceito CAPES do programa para o triênio 2010-2012.
Pesquisadores	Número de pesquisadores em um dado programa.
Dissertações de mestrado	Número de dissertações defendidas orientadas pelos pesquisadores do programa.
Teses de doutorado	Número de teses defendidas orientadas pelos pesquisadores do programa.
Supervisões de pós-doutorado	Número de pós-doutoramentos supervisionados pelos pesquisadores do programa.
Dissertações de mestrado_PP	Número de dissertações por pesquisador do programa.
Teses de doutorado_PP	Número de teses por pesquisador do programa.
Supervisões de pós-doutorado_PP	Número de pós-doutoramentos por pesquisador do programa.
Publicações - Totais	Número total de artigos completos em periódicos ou anais publicados pelo pesquisadores do programa.
Publicações - Totais_PP	Número de artigos por pesquisador do programa.
Publicações - Pontuação	Total de pontos obtidos pelas publicações dos pesquisadores do programa.
Publicações - Pontuação_PP	Total de pontos por pesquisador do programa.
Publicações Índice Restrito - Totais	Número total de publicações dos pesquisadores do programa em veículos de índice restrito (A1, A2 ou B1).
Publicações Índice Restrito - Totais_PP	Publicações dos pesquisadores do programa em veículos e índice restrito por pesquisador.
Publicações Índice Restrito - Pontuação	Total de pontos obtidos pelas publicações dos pesquisadores do programa em veículos de índice restrito.
Publicações Índice Restrito - Pontuação_PP	Total de pontos das publicações em veículos e índice restrito por pesquisador.
Média das Citações	Média das citações totais do Google Acadêmico dos pesquisador do programa.
Média das Citações_5	Média das citações do Google Acadêmico dos pesquisador do programa nos últimos cinco anos.
Média do Índice H	Média do índice H do Google Acadêmico dos pesquisador do programa.
Média do Índice H_5	Média do índice H do Google Acadêmico dos pesquisador do programa nos últimos cinco anos.
Média do Índice I10	Média do índice H do Google Acadêmico dos pesquisador do programa.
Média do Índice I10_5	Média do índice I10 do Google Acadêmico dos pesquisador do programa nos últimos cinco anos.
Mediana das Citações	Mediana das citações totais do Google Acadêmico dos pesquisador do programa.
Mediana das Citações_5	Mediana das citações do Google Acadêmico dos pesquisador do programa nos últimos cinco anos.
Mediana do Índice H	Mediana do índice H do Google Acadêmico dos pesquisador do programa.
Mediana do Índice H_5	Mediana do índice H do Google Acadêmico dos pesquisador do programa nos últimos cinco anos.
Mediana do Índice I10	Mediana do índice H do Google Acadêmico dos pesquisador do programa.
Mediana do Índice I10_5	Mediana do índice I10 do Google Acadêmico dos pesquisador do programa nos últimos cinco anos.

classe e baixa intercorrelação são preferíveis [Hall 1998]. Este seletor trabalha associado a um método de busca. Fez-se uso do método *BestFirst*, sendo que a versão utilizada

Tabela 2. Atributos de rede utilizados

Métricas Locais - coautorias entre programas	
Sigla utilizada	**Descrição**
Centralidade de grau	Medida da importância de um programa com base em seu grau na rede de coautorias.
Centralidade de intermediação	Medida da importância de um programa com base na quantidade de vezes que se encontra nos caminhos mínimos entre programas na rede de coautorias.
Centralidade de proximidade	Medida da importância de um programa com base na média de seus caminhos mínimos com os demais programas na rede de coautorias.
Centralidade Page Rank	Medida da importância de um programa com base na medida Page Rank (medida que combina as relações de coautoria com a importância dos coautores).
Métricas Globais - coautorias inter-programas	
Sigla utilizada	**Descrição**
Arestas	Número de arestas na rede de coautorias de cada programa.
Média dos caminhos mínimos	Tamanho médio dos caminhos mínimos na rede de coautorias de cada programa.
Coeficiente de agrupamento	Medida de transitividade das relações de coautoria dentro de cada programa.
Assortatividade de grau	Medida da tendência de haver relacionamentos entre autores com o mesmo grau na rede.
Diâmetro	Diâmetro (maior caminho mínimo) da rede de coautorias de cada programa.
Densidade	Densidade (relação entre o número de arestas existentes e o número de arestas possíveis) na rede de coautorias de cada programa.
Componente Gigante	Porcentagem de nós no componente gigante (maior componente conexo) da rede de cada programa.
Centralização de grau	Medida da dependência da rede de coautorias de cada programa em relação ao seu nó mais importante (de acordo com o grau).
Centralização de intermediação	Medida da dependência da rede de coautorias de cada programa em relação ao seu nó mais importante (de acordo com a intermediação).
Centralização de proximidade	Medida da dependência da rede de coautorias de cada programa em relação ao seu nó mais importante (de acordo com a proximidade).

pelo arcabouço Weka também foi proposta pelo mesmo autor [Hall 1998]. Este consiste numa busca através do espaço dos subconjuntos de atributos fazendo uso de *hillclimbing* e *backtracking*.

Para a classificação, foram testados diferentes classificadores utilizando como estratégia de verificação dos resultados a validação cruzada em dez subconjuntos. Diferentes conjuntos de atributos foram utilizados (incluindo o uso de todos os atributos ou apenas daqueles selecionados pelos seletores de atributos). Os classificadores que atingiram os melhores resultados e compõe os modelos produzidos ao final dos testes foram: *BayesNet*, *NaiveBayes* e *RandomTree* e os melhores resultados foram aqueles utilizando os atributos selecionados por *CFSSubsetEval*.

Rede Bayesiana (*BayesNet*) corresponde à aplicação do Teorema de Bayes a modelos probabilísticos representados por grafos acíclicos direcionados. A estruturação da rede pode ser feita de maneira automática ao utilizar-se de um método de busca aliado a um sistema de avaliação [Korb and Nicholson 2010]. O algoritmo *NaiveBayes*, assim como o *BayesNet*, trata-se da aplicação do Teorema de Bayes a modelos probabilísticos. Entretanto, neste algoritmo aplica-se o teorema aos atributos do conjunto de dados assumindo forte independência entre estes [John and Langley 1995]. *RandomTree* trata-se de

um algoritmo de árvore de decisão no qual são escolhidos k atributos aleatórios a fim de realizar a classificação de uma instância [Hastie et al. 2001].

A próxima seção apresenta e discute os resultados da análise dos atributos e classificação dos programas.

3. Resultados

No Brasil, o oferecimento regular de programas de pós-graduação é condicionado à obtenção de conceitos CAPES entre 3 e 7. Na avaliação atual da CAPES, um programa na área de Ciência da Computação recebeu nota 2. Assim, dividiu-se o processo de classificação em duas principais vertentes: uma fazendo uso do conjunto de dados inteiro incluindo-se o programa de nota 2 e outra incluindo apenas programas com notas de 3 a 7.

A figura 1 apresenta os valores de correlação entre cada um dos atributos extraídos ou calculados e o Conceito CAPES atual dos programas. Observa-se que os valores das correlações considerando todos os programas ou apenas aqueles que possuem conceitos entre 3 e 7 são muito próximos. Nessa figura, os atributos estão ordenados de acordo com o valor de suas correlações (do maior para o menor). Os valores elevados de correlação indicam que quanto maior o valor de um dado atributo há maior chance do programa obter um conceito CAPES maior.

O maior valor de correlação ocorre entre o Conceito CAPES atual (quadriênio 2013 a 2016) e o Conceito CAPES anterior (triênio 2010 a 2012). As três correlações seguintes possuem valores muito próximos e estão ligadas às publicações ou pontuações obtidas via publicações, são elas: Publicações Índice Restrito - Totais, Publicações - Pontuação, Publicações Índice Restrito - Pontuação. A quarta maior correlação ocorre com o atributo Teses de doutorado. A sexta e a sétima maiores correlações ocorrem com duas medidas de centralidade ao se considerar a rede em que cada programa corresponde a um nó: Centralidade de grau e Centralidade Page Rank.

Apenas uma medida diretamente relacionada às citações dos artigos dos orientadores aparece entre as dez maiores correlações: Média do Índice H_5 (isto é, a média do índice H considerando-se as citações recebidas nos últimos cinco anos). A nona maior correlação ocorre com o atributo Publicações - Totais (número total de artigos publicados em anais de eventos ou periódicos). Por outro lado, a décima maior correlação ocorre com uma medida "por pesquisador": Publicações Índice Restrito - Pontuação_PP que é a mesma medida que apresentou a quarta maior correlação, porém ponderada pelo número de pesquisadores do programa. Destaca-se que todos os atributos avaliados "por pesquisador" apresentaram correlações inferiores do que aquelas obtidas pelo valor total (não ponderado) da respectiva medida.

Apenas quatro atributos apresentaram correlações negativas, sendo que nenhuma delas possui valor absoluto alto. Destaca-se apenas a correlação com valor mais alto, que ocorreu com o atributo Densidade. Esta correlação, apesar de não possuir valor muito alto, indica que há uma relação entre a rede de coautorias do programa ser mais densa e o programa possuir um conceito CAPES menor. Isto pode sugerir que programas cujas coautorias estão muito concentradas dentro do programa acabam tendo uma produção mais limitada do que aqueles cujas coautorias ocorrem mais frequentemente com colaboradores externos ao programa.

	Todos os programas	Programas com conceitos de 3 a 7
Conceito CAPES Anterior	0.920	0.924
Publicações Índice Restrito - Totais	0.887	0.889
Publicações - Pontuação	0.886	0.888
Publicações Índice Restrito - Pontuação	0.885	0.887
Teses de doutorado	0.850	0.854
Centralidade de grau	0.821	0.818
Centralidade Page Rank	0.819	0.816
Média do Índice H_5	0.803	0.807
Publicações - Totais	0.805	0.802
Publicações Índice Restrito - Pontuação_PP	0.803	0.800
Média do Índice H	0.796	0.801
Publicações Índice Restrito – Totais_PP	0.795	0.793
Média das Citações_5	0.792	0.795
Média do Índice I10_5	0.783	0.786
Média das Citações	0.778	0.781
Mediana do Índice H_5	0.776	0.775
Média do Índice I10	0.765	0.769
Teses de doutorado_PP	0.767	0.767
Publicações - Pontuação_PP	0.763	0.762
Dissertações de mestrado	0.760	0.758
Centralidade de proximidade	0.736	0.735
Pesquisadores	0.730	0.724
Centralidade de intermediação	0.714	0.714
Mediana do Índice H	0.674	0.665
Supervisões de pós-doutorado	0.641	0.645
Arestas	0.639	0.634
Mediana das Citações	0.628	0.629
Mediana do Índice I10	0.560	0.584
Diâmetro	0.559	0.548
Média dos caminhos mínimos	0.548	0.538
Supervisões de pós-doutorado_PP	0.419	0.425
Mediana das Citações_5	0.414	0.414
Mediana do Índice I10_5	0.415	0.413
Publicações - Totais_PP	0.370	0.361
Dissertações de mestrado_PP	0.297	0.295
Componente Gigante	0.282	0.273
Assortatividade de grau	0.219	0.199
Centralização de intermediação	0.198	0.182
Coeficiente de agrupamento	-0.053	-0.036
Centralização de proximidade	-0.122	-0.128
Centralização de grau	-0.228	-0.219
Densidade	-0.245	-0.234

Figura 1. Correlações entre os atributos utilizados e o conceito CAPES atribuídos aos programas

As demais análises realizadas nesta seção foram divididas em duas subseções, uma considerando todos os programas avaliados e outra apenas para os programas com conceitos de 3 a 7.

3.1. Análise para todos os programas avaliados

A tabela 3 apresenta os dez atributos melhor ranqueados de acordo com o seletor de atributos *ChiSquaredAttributeEval*. Observa-se que seis dessas medidas (ou atributos) são relacionadas às publicações ou citações, sendo duas relacionadas às pontuações baseadas nos extratos Qualis atribuídos aos veículos nos quais os artigos foram publicados

(Publicações Índice Restrito - Pontuação e Publicações - Pontuação), à quantidade de artigos publicados (Publicações Índice Restrito - Totais e Publicações - Totais) e às citações recebidas pelos artigos dos orientadores do programa (Média do Índice H e Média do Índice I10).

Tabela 3. Atributos melhor ranqueados - método baseado no valor qui-quadrado - programas notas 2 a 7

Atributo	Valor
Conceito CAPES Anterior	113,06
Dissertações de mestrado	111,52
Publicações Índice Restrito - Totais	100,46
Publicações Índice Restrito - Pontuação	99,80
Centralidade de Grau	97,46
Publicações - Pontuação	96,17
Média do Índice H	86,95
Publicações - Totais	79,99
Teses de doutorado	79,78
Média do Índice I10	79,67

Os demais atributos melhor ranqueados são o número total de Dissertações de mestrado e de Teses de doutorado defendidas no quadriênio e Centralidade de Grau que é a única métrica oriunda da análise de redes sociais entre os dez atributos listados, ocupando a quinta posição deste ranqueamento.

Destaca-se que o atributo melhor ranqueado foi Conceito CAPES Anterior, lembrando-se que mais de dois terços dos programas mantiveram a nota obtida no triênio anterior. Este atributo é seguido pelo número total de Dissertações de mestrado defendidas e Publicações Índice Restrito - Pontuação, isto é, a pontuação total obtida pelas publicações no quadriênio em veículos classificados com Qualis A1, A2 e B1.

O algoritmo *CFSSubsetEval* com método de busca *BestFirst* tem por objetivo selecionar o subconjunto de atributos mais representativos, produzindo uma lista não ordenada. Os atributos selecionados considerando os programas com conceito de 2 a 7 são: Publicações - Pontuação_PP, Publicações Índice Restrito - Totais, Publicações Índice Restrito - Totais_PP, Publicações Índice Restrito - Pontuação, Média do Índice H, Média do Índice H_5, Centralidade de Grau, Diâmetro, Conceito CAPES Anterior.

Dos nove atributos selecionados, cinco já haviam sido selecionados pelo algoritmo *ChiSquaredAttributeEval*: Centralidade de Grau, Conceito CAPES Anterior, Média do Índice H, Publicações Índice Restrito - Pontuação e Publicações Índice Restrito - Totais. Os atributos que não haviam sido selecionados são Diâmetro (da rede de coautorias interna de cada programa), Média do Índice H_5, Publicações - Pontuação_PP e Publicações Índice Restrito - Totais_PP. Destaca-se que este algoritmo visa a identificar um subconjunto de atributos que seja mais significativo em relação ao Conceito CAPES atual (analisados de maneira conjunta), o que é diferente de se observar os resultados anteriormente apresentados que são focados nos atributos de maneira individual.

Observa-se que nos resultados do algoritmo *CFSSubsetEval* houve preferência do uso das duas médias do índice H (considerando todas as citações e apenas as citações

dos últimos cinco anos), bem como da escolha de uma medida de rede que não havia sido selecionada anteriormente (Diâmetro) e também de duas medidas "por pesquisador": Publicações - Pontuação_PP e Publicações Índice Restrito - Totais_PP. Isto revela que estes atributos, apesar de não apresentarem as maiores correlações com o Conceito CAPES dos programas, são relevantes para a atribuição desses conceitos.

Ao aplicar ao subconjunto obtido pelo algoritmo *CFSSubsetEval* o algoritmo de classificação *NaiveBayes* foi possível classificar corretamente 81,82% das instâncias. É importante salientar também que o erro de classificação em todos os casos foi de apenas um ponto no conceito (por exemplo, todos programas de nota 4 foram classificados como programas de conceito 3, 4 ou 5), conforme pode ser observado na matriz de confusão (tabela 4), na qual a primeira linha representa as classificações do modelo e a última coluna apresenta o valor real do conceito daquele programa.

Tabela 4. Matriz de Confusão - classificação de todos os programas avaliados

a	b	c	d	e	f	
0	1	0	0	0	0	a = 2
0	21	3	0	0	0	b = 3
0	3	20	1	0	0	c = 4
0	0	2	5	0	0	d = 5
0	0	0	0	2	1	e = 6
0	0	0	0	1	6	f = 7

Não havendo conjunto de treinamento a parte para a criação do modelo, foi utilizado o método de validação cruzada em 10 *folds* para o treinamento e validação do modelo gerado. Vale salientar que, por haver apenas um programa de nota 2, quando este era utilizado no subconjunto de treinamento, nenhuma instância dessa classe seria utilizada como teste. Com isso, seria impossível para o algoritmo classificar corretamente este programa.

3.2. Análise para os programas avaliados com conceitos de 3 a 7

A tabela 5 apresenta os dez atributos melhor ranqueados de acordo com o seletor de atributos *ChiSquaredAttributeEval*. Observa-se que os atributos selecionados são os mesmos presentes na tabela 3, ocorrendo apenas mudanças na ordem de alguns atributos. Os quatro primeiros atributos aparecem em ordem diferente daqueles da tabela 3. Estas mudanças ocorreram porque, ao excluir o programa com conceito 2 da análise, o atributo Publicações Índice Restrito - Pontuação subiu da quarta posição para a primeira, deslocando os três atributos que haviam sido melhor ranqueados do que ele.

Destaca-se, assim, que neste conjunto de dados o atributo melhor ranqueado foi Publicações Índice Restrito - Pontuação, sendo melhor ranqueado do que o Conceito CAPES Anterior. Isto indica a importância desta medida que combina a publicação de artigos com a "qualidade" dos veículos nos quais os artigos foram publicados considerando os extratos Qualis A1, A2 e B1. Observa-se ainda que este atributo considera a pontuação total do programa e não aquela dividida pelo número de orientadores.

Fazendo uso do seletor de atributos *CFSSubsetEval* com o algoritmo de busca *BestFirst* foi possível reduzir o número de atributos a 10, sendo estes: Centralidade de

Tabela 5. Atributos melhor ranqueados - método baseado no valor qui-quadrado - todos os programas

Atributo	Valor
Publicações Índice Restrito - Pontuação	112,91
Conceito CAPES Anterior	110,65
Dissertações de mestrado	109,72
Publicações Índice Restrito - Totais	98,64
Centralidade de Grau	95,34
Publicações - Pontuação	94,38
Média do Índice H	84,32
Publicações - Totais	78,45
Teses de doutorado	78,36
Média do Índice I10	77,94

Grau, Conceito CAPES Anterior, Diâmetro, Dissertações de mestrado Média das Citações Média do Índice H, Publicações - Pontuação_PP, Publicações Índice Restrito - Pontuação, Publicações Índice Restrito - Totais, Publicações Índice Restrito - Totais_PP.

Destacam-se as diferenças ocorridas na seleção atual e naquele que envolvia todos os programas. Na seleção atual, não consta o atributo Média do Índice H_5, o qual havia sido selecionado ao se considerar todos os programas. Por outro lado, ao se considerar apenas os programas com conceitos de 3 a 7, dois novos atributos foram selecionados: Dissertações de mestrado (número de dissertações defendidas no quadriênio) e Média das Citações (média das citações recebidas pelos orientadores ao longo de suas carreiras).

O melhor resultado da classificação considerando-se os programas com conceitos de 3 a 7 ocorreu utilizando-se o conjunto de atributos selecionados pelo *CFSSubsetEval*. Foi realizada uma classificação em dois níveis. Inicialmente separou-se os programas de nota 6 e 7 dos demais e depois classificaram-se os demais programas. O classificador *RandomTree* foi capaz de classificar com 100% de acerto todos programas com conceitos 6 e 7, sem nenhum falso-positivo. Desta forma, é possível utilizar o resultado deste classificador para separar os programas com conceitos 6 e 7 e então usar uma nova abordagem para classificar os programas restantes. A tabela 6 apresenta a matriz de confusão referente a esta classificação.

Tabela 6. Matriz de Confusão - primeiro nível de classificação dos programas com conceitos de 3 a 7

a	b	c	d	e	
16	7	1	0	0	a = 3
6	14	4	0	0	b = 4
0	4	3	0	0	c = 5
0	0	0	3	0	d = 6
0	0	0	0	7	e = 7

Para os programas com conceitos entre 3 e 5 foi realizada uma nova seleção de atributos (utilizando a mesma técnica citada anteriormente) que resultou no seguinte sub-

conjunto: Centralidade de Grau, Conceito CAPES Anterior, Dissertações de mestrado Média das Citações_5 Média do Índice H, Média dos caminhos mínimos, Publicações - Pontuação_PP, Publicações Índice Restrito - Pontuação, Publicações Índice Restrito - Totais, Publicações Índice Restrito - Totais_PP.

Esta seleção de atributos difere da anterior em dois atributos. O atributo Média das Citações_5 está substituindo o atributo Média das Citações que havia sido selecionado anteriormente, indicando que para o conjunto de programas com conceitos de 3 a 5 a média das citações recebidas nos últimos cinco anos é mais importante do que a média de todas as citações recebidas. A outra diferença ocorre com o atributo Média dos caminhos mínimos que foi selecionado no lugar do Diâmetro.

Aplicando o classificador *BayesNet* a esse novo conjunto (considerando-se apenas os programas com conceitos de 3 a 5) obteve-se índice de acerto de 83,64%. A tabela 7 apresenta a matriz de confusão referente a esta classificação.

Tabela 7. Matriz de Confusão - programas com conceitos de 3 a 5

a	b	c	
7	0	0	a = 3
1	18	5	b = 4
0	3	21	c = 5

Com esta classificação em duas etapas, ao analisar os programas de nota 3 a 7, obteve-se 56 classificações corretas num total de 65 programas resultando num modelo com acurácia de 86,15%.

4. Conclusões

Neste trabalho foram medidos e analisados diferentes atributos ou características dos programas brasileiros de pós-graduação em Ciência da Computação do Brasil. Analisou-se a importância desses atributos em relação aos conceitos atribuídos pela CAPES em sua avaliação quadrienal, bem como a capacidade de se inferir o conceito com base nestes atributos.

Observando-se os resultados alcançados, é possível verificar que, dentre os métodos analisados, obteve-se melhores resultados ao se utilizar classificadores Bayesianos sobre subconjuntos de atributos selecionados tendo como objetivo a minimização da correlação entre estes e a maximização da correlação de cada atributo com o atributo classe, visando assim a eliminação de redundâncias.

Com o modelo obtido é possível, por exemplo, automatizar parte do processo de avaliação destes programas, bem como utilizá-lo para auto-avaliação por parte das universidades responsáveis pelos programas em janelas de tempo diferentes daquela em uso pela CAPES.

Como trabalhos futuros pretende-se explorar outras estratégias de seleção de atributos e classificação, bem como estender a análise realizada para programas avaliados por outros comitês.

Agradecimentos

Este trabalho foi parcialmente financiado pelo Programa de Educação Tutorial (PET) do Ministério da Educação e pelo CNPq.

Referências

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. A. and Ferreira, J. E. (2018). Desambiguação de nomes de autores para a identificação automática de perfis acadêmicos. *Em Questão*, pages 1–12.

Digiampietri, L. A., Mena-Chalco, J., de Jesus Perez-Alcazar, J., Tuesta, E. F., Delgado, K., and Mugnaini, R. (2012). Minerando e caracterizando dados de currículos Lattes. In *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2012)*.

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. (2014). Brax-ray: An x-ray of the brazilian computer science graduate programs. *PLOS ONE*, 9(4):1–12.

Hall, M. A. (1998). *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, University of Waikato, Hamilton, New Zealand.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

John, G. H. and Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, San Mateo. Morgan Kaufmann.

Korb, K. B. and Nicholson, A. E. (2010). *Bayesian Artificial Intelligence, Second Edition*. CRC Press, Inc., Boca Raton, FL, USA, 2nd edition.

Linden, R., Barbosa, L. F., and Digiampietri, L. A. (2017). "Brazilian style science" – an analysis of the difference between Brazilian and international computer science departments and graduate programs using social networks analysis and bibliometrics. *Social Network Analysis and Mining*, 7(1):44.

Mena-Chalco, J. P., Digiampietri, L. A., Lopes, F. M., and Cesar, R. M. (2014). Brazilian bibliometric coauthorship networks. *Journal of the Association for Information Science and Technology*, 65:1424–1445.

Silva, T. H. P., Laender, A. H. F., Davis, C. A., da Silva, A. P. C., and Moro, M. M. (2017). A profile analysis of the top Brazilian computer science graduate programs. *Scientometrics*, 113(1):237–255.

Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining, Fourth Edition: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 4th edition.