WILLIAM TAKAHIRO MARUYAMA

Predição de coautorias em redes sociais acadêmicas

São Paulo

2016

WILLIAM TAKAHIRO MARUYAMA

Predição de coautorias em redes sociais acadêmicas

Dissertação apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação.

Área de concentração: Metodologia e Técnicas da Computação

Versão corrigida contendo as alterações solicitadas pela comissão julgadora em 28 de Março de 2016. A versão original encontra-se em acervo reservado na Biblioteca da EACH-USP e na Biblioteca Digital de Teses e Dissertações da USP (BDTD), de acordo com a Resolução CoPGr 6018, de 13 de outubro de 2011.

Orientador: Prof. Dr. Luciano Antonio Digiampietri

São Paulo

2016

CATALOGAÇÃO-NA-PUBLICAÇÃO
(Universidade de São Paulo. Escola de Artes, Ciências e Humanidades. Biblioteca)

Maruyama, William Takahiro
Predição de coautorias em redes sociais acadêmicas / William Takahiro Maruyama ; orientador, Luciano Antonio Digiampietri. – São Paulo, 2016
154 f. : il.

Dissertação (Mestrado em Ciências) - Programa de Pós-Graduação em Sistemas de Informação, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo
Versão corrigida

1. Tecnologia da informação. 2. Ciência da computação. 3. Redes sociais - Análise. 4. Pesquisadores. I. Digiampietri, Luciano Antonio, orient. II. Título

CDD 22.ed.- 004

Dissertação de autoria de William Takahiro Maruyama, sob o título **"Predição de coautorias em redes sociais acadêmicas"**, apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação, na área de concentração Metodologia e Técnicas da Computação, aprovada em 28 de Março de 2016 pela comissão julgadora constituída pelos doutores:

Prof. Dr. Luciano Antonio Digiampietri
Presidente
Universidade de São Paulo

Prof. Dr. Pedro Olmo Stancioli Vaz De Melo
Universidade Federal de Minas Gerais

Profa. Dra. Karina Valdivia Delgado
Universidade de São Paulo

Aos meus pais.

Agradecimentos

Em primeiro lugar, agradeço a minha família - minha mãe Satiko, meu pai Nelson e minha irmã Mayara - por todo apoio incondicional que me possibilitaram chegar até aqui.

Agradecimento especial ao meu orientador Prof. Dr. Luciano Antonio Digiampietri pela parceria, conhecimento passado, dedicação e paciência durante todo o mestrado.

À minha namorada Natalia pelo apoio, paciência e companheirismo. Além das revisões dos meus textos e seus comentários valiosos.

Aos professores e ex-professores da EACH que contribuíram para minha formação acadêmica ao longo da graduação e do mestrado.

Aos meus amigos pelo apoio e incentivo para seguir em frente.

Aos meus colegas do PPgSI que me acompanharam e que me ajudaram durante esse período.

Por fim, agradeço a CAPES e a Universidade de São Paulo.

Resumo

MARUYAMA, William Takahiro. **Predição de coautorias em redes sociais acadêmicas**. 2016. 154 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2016.

Atualmente, as redes sociais estão ganhando cada vez mais destaque no dia-a-dia das pessoas. Nessas redes são estabelecidos diferentes relacionamentos entre entidades que compartilham alguma característica ou objetivo em comum. Diversas informações sobre a produção científica nacional podem ser encontradas na Plataforma Lattes, que é um sistema utilizado para o registro dos currículos dos pesquisadores no Brasil. A partir dessas informações é possível construir uma rede social acadêmica, na qual as relações entre os pesquisadores representam uma parceria na produção de uma publicação (coautoria) - um *link*. Na análise de redes sociais existe uma linha de pesquisa conhecida como predição de *links* ou de relacionamentos, que tem como objetivo identificar relacionamentos futuros. Essa tarefa pode favorecer a comunicação entre os usuários e otimizar o processo de produção científica identificando possíveis colaboradores. Este projeto analisou a influência de diferentes atributos encontrados na literatura e filtros de dados para prever relações de coautoria nas redes sociais acadêmicas. Foi abordado dois tipos de problemas na predição de relacionamentos, o problema geral que analisa todos os possíveis relacionamentos de coautoria e o problema de novas coautoria que refere-se aos relacionamentos de coautorias inéditas na rede. Os resultados dos experimentos foram promissores para o problema geral de predição com a combinação de atributos e filtros utilizados. Contudo, para o problema de novas coautorias, devido à sua maior complexidade, os resultados não foram tão bons. Os experimentos apresentados avaliaram diferentes estratégias e analisaram o custo e benefício de cada uma. Conclui-se que para lidar com o problema de predição de coautorias em redes sociais acadêmicas é necessário analisar as vantagens e desvantagens entre as estratégias, encontrando um equilíbrio entre a revocação da classe positiva e a acurácia geral.

Palavras-chaves: Predição de Links, Predição de Coautorias, Redes de Coautoria, Redes Acadêmicas, Análise de Redes Sociais.

Abstract

MARUYAMA, William Takahiro. **Link Prediction in academic social networks.** 2016. 154 p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2016.

Nowadays, social networks are gaining prominence in the day-to-day lives. In these networks, different relationships are established between entities that share some characteristic or common goal. A huge amount of information about the Brazilian national scientific production can be found in the Lattes Platform, which is a system used to record the curricula of researchers in Brazil. From this information, it is possible to build an academic social network, where relations between researchers represent a partnership in the production of a publication - a link. In social network analysis there is a research area known as link prediction, which aims to identify future relationships. This task may facilitate communication among researchers and optimize the scientific production process identifying possible collaborators. This project analyzed the influence of different attributes found in the literature and data filters to predict co-authorship relationships in academic social networks. Was approached two types of problems in predicting relationships, the general problem that analyzes all possible co-authoring relationships and the problem of new co-authoring that relates to novel co-authorships relationships in the network. The experimental results were promising to the prediction general problem, combining attributes and using filters. However, for the new co-authorships problem the results were not as good. The experiments evaluated different strategies and analyzed the costs and benefits of each. We concluded that to deal with the co-authorships prediction problem in academic social networking it is necessary to analyze the advantages and disadvantages among the strategies, finding a balance between the recall of the positive class and the overall accuracy.

Keywords: Link Prediction, Co-authorship Prediction, Coauthoring Networks, Academic Networks, Social Network Analysis.

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

AA_e	Adamic-Adar baseado em evento
AA	Adamic-Adar
ACC	Acurácia
ACM	*Association for Computing Machinery*
AD	Árvore de Decisão
ADTree	*Alternating Decision Tree*
AF	*Affinity measure* (CHANG; YAO, 2011)
AF	*Average Filling* (HUANG et al., 2012)
aKatz	*Approximate Katz*
AL	*Absent Links*
AL	*Attention Limited*
APG	*Accelerated Proximal Gradient*
API	Interface de Programação de Aplicação
Astro-ph	*Astrophysics*
AT	*Attractiveness*
AT-PRP	*Attractiveness com PageRank*
AUC	*Area Under the Curve*
Av	*Average*
BC	*Betweenness centrality*
BFTree	*Best-First Decision Tree*
BH-CRM	*Bayesian Hierarchical Community-and-Role Model*
BOW	*Bag of Words*

BPG	*Bootstrap Probabilistic Graph*
BrAA	*BenefitRanked Adamic-Adar*
BrCN	*BenefitRanked Commom Neighbor*
BrRA	*BenefitRanked Resource Allocation*
CA_e	*Common attendees*
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CART	*Classification and Regression Trees*
CDR	*Call Detail Record* / Registro de Detalhes da Chamada
CF_i	*Item-based collaborative filtering*
CF_u	*User-based collaborative filtering*
CLRA-CN	*Clustered Low Rank Approximation with Commom Neighbor*
CLRA-Katz	*Clustered Low Rank Approximation with Katz*
CN	*Common Neighbor*
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
Cond-mat	*Condensed Matter*
CP	*Category Popularity*
CS	*Conservative Metric*
CS_AL	*Attention-limited Conservative Metric*
CSGE	*Clustered Spectral Graph Embedding*
DBLP	*Digital Bibliography & Library Project*
dIRM	*dynamic Infinite Relational Model*
dMMSB	*dynamic Mixed Membership Stochastic Blockmodels*
DMNB	*Discriminative Multinominal Naive Bayes*

dRTM	*dynamic Relational Topic Model*
DRW	*Weighted call duration random walk*
DT	*Decision Table* / Tabela de Decisão
DTNB	*Decision Rable Naive Bayes*
E	*Ethnicity*
EIG-CN	*Eigen Decomposition with Commom Neighbor*
EIG-Katz	*Eigen Decomposition with Katz*
ERGM	*Exponential random graph model*
FN	Falso Negativo
FP	Falso Positivo
FPC	*Fixed Point Continuation*
FT	*Functional Trees*
GARSC	Grupo de Análise de Redes Sociais e Cientometria
GC	*Common Groups*
GD	*Graph Distance*
GEFR	*Geo-Friends Recommendation Framework*
GJC	*Jaccard's Coefficient for Groups*
GLFM	*Generalized Latent Factor Model*
GNMF	*Graph Nonnegative Matrix Factorization*
GPS	*Global Positioning System* / Sistema de Posicionamento Global
GPSSim	*Global Positioning System Similarity*
GRJMF	*Graph Regularized Joint Matrix Factorization*
Gr-qc	*General relativity and quantum cosmology*

GSBM	*Generalized Stochastic Blockmodel*
HDI	*Hub Depressed Index*
Hep-lat	*High energy physics lattice*
Hep-ph	*High energy physics phenomenology*
Hep-th	*High energy physics theory*
HITS	*Hyperlink-Induced Topic Search*
HPI	*Hub Promoted Index*
HPLP	*High-Performance Link Prediction*
IC	*Common Interests*
IEEE	*Institute of Electrical and Electronics Engineers*
IJC	*Jaccard's Coefficient for Interests*
IO	*Item Ownership*
ITCom	*Time-evolving Composite Network Models*
JC	*Jaccard's Coefficient*
JMF	*Joint Manifold Factorization*
JS	*Jaccard Similarity*
Katz-C	*Katz based on all sources*
Katz-S	*Katz based on a single source*
KDD	*Knowledge Discovery in Databases*
LDA	*Latent Dirichlet Allocation*
LDA-G	*Latent Dirichlet Allocation for Graphs*
LES	*Linear Exponential Smoothing*
LFBM	*Latent Factor BlockModel*

LHN *Leicht-Holme-Newman Index*

LP *Local Path*

LR *Linear Regression*

LSI *Latent Semantic Indexing*

LWL *Locally Weighted Learning*

Ma *Moving Average*

MC *Matrix Completion*

MDL *Minimum Description Length*

ME *Maximum Entropy*

MIT *Massachusetts Institute of Technology*

MLFM *Multiplicative Latent Factor Model*

MLP *Multi-Layer Perceptron* / Perceptron Multicamadas

MMB *Mixed Membership Stochastic Blockmodel*

MMSB *Mixed Membership Stochastic Blockmodels*

MSLP-CN *Multi-Scale Link Prediction with Commom Neighbors*

MSLP-Katz *Multi-Scale Link Prediction with Katz*

MWF *Merge Weighted Features*

NB *Naive Bayes*

NBTree *Naive Bayes Tree*

NC *Non-Conservative Proximity*

NC_AL *Attention-Limited Non-Conservating Proximity*

NMDR *Nonparametric Metadata Dependent Relational Model*

NMF *Nonnegative Matrix Factorization*

PA	*Preferential Attachment*
PAM	*Professional Activity Match*
PB	*Political blogs network*
PC	Componentes Principais
PCA	*Principal Component Analysis* / Análise de Componentes Principais
PD	*Path Distance*
PG	*Power grid Network*
PR	*PageRank*
PRP	*PageRank with Priors* (KUO et al., 2013)
PRP	*Page Rank Product* (SONG et al., 2009)
RA	*Resource Allocation Index*
RADRW	*Resource allocation based on weighted call duration random walk*
RATRW	*Resource allocation based on weighted call times random walk*
RAURW	*Resource allocation based on weighted random walk*
RBF	*Radial Basis Function* / Função de Base Radial
RC	*Common Regions*
REP	*Reduced-Error Pruning*
REPTree	*Reduced-Error Pruning Tree*
RFG	*Ranking Factor Graph model*
RIPPER	*Repeated Incremental Pruning to Produce Error Reduction*
RO	*Observations Together*
ROC	*Receiver Operating Characteristic*
RPR	*Rooted Pagerank*

RS *Regions Seen Concurrently*

RTM *Relational Topic Model*

RW *Random Walk*

RWR *Random Walk with Restart*

SA *Salton Index*

SES *Simple Exponential Smoothing*

SGE *Spectral Graph Embedding*

SIAM *Society of Industrial and Applied Mathematics publications*

SL-H *Supervised learning with hybrid color paths*

SL-H(HS) *Supervised learning with hybrid color paths (hierarchical structured regularization)*

SL-H(L1) *Supervised learning with hybrid color paths (L1 regularization)*

SL-P *Supervised learning with only pure color paths*

SL-P(L1) *Supervised learning with only pure color paths (L1 regularization)*

SL-S *Supervised learning with single source*

SMO *Sequential Minimal Optimization* / Otimização Mínima Sequencial

SNA / ARS *Social Network Analysis* / Análise de Redes Sociais

SO *Sorensen Index*

SofN *Sum of Neighbors*

SofP *Sum of Patients*

SP *Shortest Path* / Menor distância

SS_Uniform *Uniform Weighting Scheme*

SVD *Singular value decomposition*

SVM	*Support Vector Machine* / Máquina de Vetores de Suporte
TF	*Tensor Factor*
TIG	*TakingltGlobal*
tKatz	*truncated Katz*
tKatz-C	*truncated Katz based on all sources*
tKatz-CS	*Truncated Katz with common subspace*
tKatz-LFM	*Truncated Katz with latent factor model*
tKatz-LFM-c	*Truncated Katz with clustered latent factor model*
tKatz-S	*truncated Katz based on single source*
TME	*Time-Aware Maximum Entropy*
TRFG	*Transfer-based Factor Graph model*
TRW	*Weighted call times random walk*
UF	*User friendship*
URW	*Unweighted Random Walk*
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo
WAA	*Weighted Adamic-Adar*
WCN	*Weighted Commom Neighbor*
Weka	*Waikato Environment for Knowledge Analysis*
WIC	*Within And Inter Cluster*
WRA	*Weighted Resource Allocation*

Sumário

[1] De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

1 Introdução

O ser humano constitui e vive em diferentes grupos de indivíduos. As interações sociais, intra e inter grupos, permitem às pessoas estabelecerem diferentes tipos de relacionamentos ao longo da vida (identificação, amizade, familiar, profissional, etc). Estes relacionamentos criam vínculos ou ligações muitas vezes intangíveis entre as pessoas, formando grupos de diferentes tipos e tamanhos. Essa organização social possibilita, em muitos casos, a cooperação para atingirem um objetivo em comum.

As redes sociais são modelos de interação entre entidades (HASAN; ZAKI, 2011). Essas entidades podem ser pessoas ou organizações, as quais compõem uma estrutura de elementos conectados entre si. As conexões dessa estrutura podem ser estabelecidas a partir de diversos tipos de relações, que podem representar algum tipo de identificação, interação, interesse, colaboração ou influência entre as entidades (LIBEN-NOWELL; KLEINBERG, 2003).

Para fomentar a interação entre as pessoas, a distância geográfica sempre foi um fator limitante. Entretanto, com os avanços tecnológicos que impulsionaram a globalização, o surgimento e a evolução da Web, muitos serviços e ferramentas foram criados ao longo dos anos para diminuir, virtualmente, a distância entre as pessoas. Essas mudanças estimularam, principalmente, a troca, a produção e a divulgação de informações das mais variadas fontes. Um exemplo de serviço popular na era digital são as redes sociais *online*, que visam possibilitar ou facilitar as relações sociais em um ambiente virtual como o Facebook[1], o Instragram[2], o LinkedIn[3], entre outras.

Dentre as diversas atividades relacionadas ao estudo das redes sociais está a predição de relacionamentos (ou predição de *links*), que consiste na identificação de possíveis relacionamentos futuros dentro de redes sociais. A predição de novos relacionamentos dentro de uma rede social é uma tarefa que ganhou bastante destaque nos últimos anos, pois pode ser utilizada para encontrar desde amigos que ainda não estavam ligados em uma rede social *online* (QUERCIA; CAPRA, 2009; VASUKI et al., 2010; TIAN et al., 2010; FIRE et al., 2011; PEREZ; BIRREGAH; LEMERCIER, 2012; ZHONG et al., 2013), até potencializar a realização de trabalhos em empresas ou na comunidade científica (SA; PRUDENCIO, 2011; DONG et al., 2012; HSIEH et al., 2013). Contudo, a predição de relacionamentos é um problema complexo devido ao comportamento altamente dinâmico da estrutura de

[1] https://www.facebook.com/
[2] http://instagram.com
[3] http://www.linkedin.com/

uma rede social, a qual se altera rapidamente ao longo do tempo e a cada nova iteração (LIBEN-NOWELL; KLEINBERG, 2003).

1.1 Motivação e Justificativa

Cada vez mais a pesquisa científica está se transformando em uma atividade colaborativa e, muitas vezes, multidisciplinar. A formação de equipes adequadas bem como a identificação das expertises necessárias são desafios complexos e necessários no processo da produção científica. Deste modo, o presente trabalho explorou um tipo específico de redes sociais: as redes sociais acadêmicas e, em particular, as redes sociais formadas pelos relacionamentos de coautoria na produção de artigos científicos.

Uma rede social é comumente representada por meio de um grafo, logo, na rede de coautoria acadêmica, os pesquisadores são representados pelos nós e as relações de coautoria pelas arestas (NEWMAN, 2010). Nas redes sociais acadêmicas, a predição de relacionamentos tem sido utilizada principalmente para a predição de coautorias (GUO; GUO, 2010; MAKREHCHI, 2011; DONG et al., 2011a; LIN; YUN; ZHU, 2012; GAO; DENOYER; GALLINARI, 2012; DIGIAMPIETRI; SANTIAGO; ALVES, 2013), atividade que indica se um par de pesquisadores poderá/deverá colaborar na produção de um artigo, podendo assim otimizar a produção destes pesquisadores por meio da indicação de pesquisadores cujas parcerias são mais promissoras. Assim, esse tipo de predição pode ser utilizado para favorecer a comunicação entre os pesquisadores por meio da sugestão de possíveis relacionamentos (*links*), almejando potencializar o processo de produção científica.

Neste contexto, o problema de predição de coautorias pode ser dividido em predição de coautorias novas/inéditas (isto é, prever quais pares de pesquisadores que nunca colaboraram na publicação de um artigo irão colaborar) e em predição geral de coautorias (predizer quais pares de pesquisadores irão colaborar na publicação de um ou mais artigos independentemente deles já terem ou não colaborado).

Alguns dos fatores que tornam a tarefa de predição de relacionamentos complexa são: a identificação de quais atributos individuais (relacionados, por exemplo, ao perfil ou currículo das pessoas) serão utilizados; especificação ou seleção de métricas estruturais de redes sociais a serem usadas; utilização de estratégias para combinar estes atributos de forma a possibilitar a predição; e o fato do conjunto de dados ser tipicamente desbalanceado

e esparso (RATTIGAN; JENSEN, 2005; CUKIERSKI; HAMNER; YANG, 2011; GAO; DENOYER; GALLINARI, 2011; HASAN; ZAKI, 2011; KUO et al., 2013).

1.2 Objetivos

Esta dissertação teve por objetivo auxiliar a enfrentar parte dos desafios da predição de relacionamentos, em particular, da predição de coautorias em redes sociais acadêmicas. Assim, o objetivo geral foi desenvolver uma solução para a predição de relacionamentos de coautoria que considere a combinação de diferentes atributos e filtros sobre os dados para prever relações de coautorias (tanto a reincidência de relações quanto relações novas/inéditas).

1.2.1 Objetivos específicos

Esta dissertação teve os seguintes objetivos específicos:

- Identificar quais atributos (características ou métricas) podem ser utilizados para a predição de relacionamentos em redes sociais;
- Identificar atributos específicos das redes de coautoria a serem utilizados;
- Analisar a influência individual do uso de cada um dos atributos ou filtros na predição de relações de coautoria;
- Desenvolver uma solução combinando os diferentes atributos e usando o filtro selecionado;
- Testar e validar a solução proposta considerando o problema geral da predição de coautorias e o problema específico de predição de relações inéditas de coautoria.

1.3 Metodologia

Para a concretização dos objetivos apresentados, a seguir será sumarizada a metodologia empregada. Nos capítulos apropriados serão explicados com mais detalhes as atividades executadas.

O delineamento do trabalho consistiu inicialmente na realização de uma revisão da literatura correlata, por meio da metodologia de revisão sistemática. Este é um método

de revisão que define e registra um processo sistemático que visa a identificar, avaliar e sumarizar os trabalhos relevantes do levantamento bibliográfico. O estudo teve como objetivo identificar quais os atributos estão sendo utilizados na predição de relacionamentos em redes sociais. Com isso, foi possível identificar uma variedade de atributos utilizados em diferentes tipos de redes sociais. Parte dos atributos encontrados foi utilizada para a consecução do presente trabalho.

Quanto à amostra, foi utilizada a versão atualizada do conjunto de dados disponibilizados pelos autores Digiampietri, Santiago e Alves (2013). Este conjunto é composto pelos 657 docentes permanentes dos programas de pós-graduação em Ciência da Computação com doutorado e/ou mestrado acadêmico que atuaram nos dois triênios 2004-2006 e 2007-2009. O conjunto de dados foi separado em treinamento e teste de acordo com as janelas de tempo. Sendo que os dados até o ano de 2010 foram utilizados para o treinamento e os modelos gerados foram testados com dados de 2011 a 2015.

A implementação da solução foi realiza em JAVA e pode ser dividida nas seguintes etapas: extração e cálculo dos atributos, pré-processamento, processamento e armazenamento dos resultados. Para algumas funções da solução, principalmente o processamento, utilizou-se a API do arcabouço do Weka.

O desempenho de cada experimento foi medido com a utilização de algumas métricas para avaliar cada estratégia adotada.

1.4 Organização da dissertação

A dissertação, incluindo este capítulo introdutório, possui seis capítulos. A seguir, é apresentada uma descrição da organização dos demais capítulos.

- O capítulo 2 apresenta, de maneira breve, os principais conceitos que envolvem o presente trabalho;
- No capítulo 3 encontra-se a revisão da literatura correlata da área de predição de relacionamentos;
- O capítulo 4 contém a metodologia utilizada no desenvolvimento da solução;
- O capítulo 5 apresenta o conjunto de dados utilizado e os resultados dos cenários de teste. Além disso, contém uma discussão sobre os resultados obtidos;

- Por fim, no capítulo 6 são apresentadas as conclusões finais, sendo abordadas as principais contribuições do trabalho e suas possíveis extensões.

2 Conceitos Fundamentais

Este capítulo apresenta brevemente os assuntos base que permeiam o presente trabalho. Nas próximas seções são apresentados: o grafo para a representação das redes (seção 2.1), a análise de redes sociais (seção 2.2) e a predição de relacionamentos (seção 2.3).

2.1 Grafos

Muitos objetos de estudo dos cientistas podem ser representados por redes. Uma rede corresponde a um conjunto de elementos ligados entre si, ou seja, redes são compostas por indivíduos ou componentes ligados conforme alguma relação entre eles (NEWMAN, 2010).

Com origem no campo da Matemática, os grafos são frequentemente utilizados para modelar redes. Isso porque a utilização da teoria dos grafos permite o uso de definições que podem ser associadas às propriedades estruturais da rede. Além disso, possibilita realizar operações matemáticas para quantificar as propriedades estruturais e também para efetuar provas matemáticas. Nos grafos, as entidades são chamadas de vértices (ou nós) e os relacionamentos são chamados de arestas (WASSERMAN; FAUST, 1994; PRELL, 2011). No presente trabalho, os nós correspondem a pessoas e as arestas representam aos relacionamentos de coautoria entre estas pessoas, isto é, uma aresta existirá para indicar que duas pessoas são coautoras de uma mesma publicação.

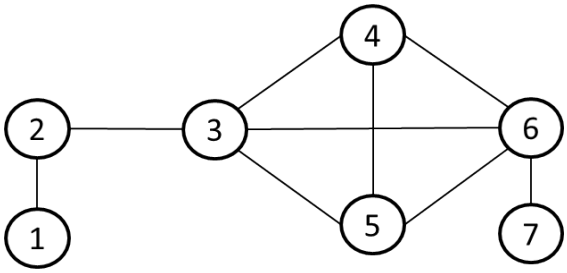
Na notação matemática, um grafo é definido como $G = (V, E)$, sendo $V = \{v_1, v_2, v_3, ..., v_m\}$ o conjunto de vértices de tamanho $m = |V|$ e $E = \{e_1, e_2, e_3, \ldots, e_n\}$, o conjunto de arestas de tamanho $n = |E|$ (CORMEN; LEISERSON; STEIN, 2012). Além disso, as arestas de um grafo podem possuir duas propriedades:

- Valores (ou pesos): podem ser binários com o significado de presença ou ausência do relacionamento (LIU et al., 2005). Além disso, os valores podem ser usados como pesos para ponderação dos relacionamentos. O valor do peso é uma forma de representar a intensidade da relação (WASSERMAN; FAUST, 1994; LIU et al., 2005; PRELL, 2011). Deste modo, além dos conjuntos V e E que formam o grafo, tem-se o conjunto $W = \{w_1, w_2, w_3, \ldots, w_n\}$ que representa os pesos das arestas. Portanto, um grafo ponderado pode ser caracterizado como $G = (V, E, W)$ (LIU et al., 2005).

- Direção: um grafo é chamado dígrafo quando as arestas são direcionadas, ou seja, as arestas possuem origem e destino, e são representadas por setas para indicar a direção. Quando um grafo é não direcionado, isto é, as arestas não possuem direção, as relações são simétricas e são representadas por linhas. As redes de coautoria são exemplos de grafos não direcionados (WASSERMAN; FAUST, 1994; LIU et al., 2005; PRELL, 2011).

A Figura 1 apresenta um exemplo da utilização de um grafo para representar as relações de coautoria. Como os relacionamentos de coautoria são simétricos, o grafo é não direcionado. Nesse grafo, os relacionamentos não são ponderados, entretanto poderiam ser, por exemplo, pela quantidade de coautorias em diferentes publicações entre os pesquisadores.

Figura 1 – Exemplo de uma rede de coautoria representada por um grafo. Os vértices representam os autores e as arestas a coautoria em pelo menos um artigo.



Artigos	Autores
Artigo A	Autor 3, Autor 4, Autor 5, Autor 6
Artigo B	Autor 4, Autor 5, Autor 6
Artigo C	Autor 6, Autor 7
Artigo D	Autor 2, Autor 3
Artigo E	Autor 1, Autor 2

Fonte: William T. Maruyama, 2015.

Os grafos são úteis para a representação visual ao criar uma abstração do relacionamento entre entidades. É possível perceber determinados padrões estruturais na rede apenas analisando-os visualmente. No entanto, isso é viável quando existe um número pequeno de vértices, pois, caso contrário, a complexidade estrutural seria muito alta para ser analisada desta forma (WASSERMAN; FAUST, 1994; NEWMAN, 2010).

Existem dois tipos de representações baseadas em estruturas de dados que são adequadas para o processamento computacional de grafos (WASSERMAN; FAUST, 1994; PRELL, 2011; CORMEN; LEISERSON; STEIN, 2012):

- Matriz de adjacência (Figura 3(a)): os vértices são representados pelas linhas e colunas de uma matriz, enquanto a existência das arestas é indicada pelo valor contido entre o cruzamento da linha e da coluna da matriz. Os valores presentes nas células podem ser binários, onde 1 refere-se a existência e 0 a não existência

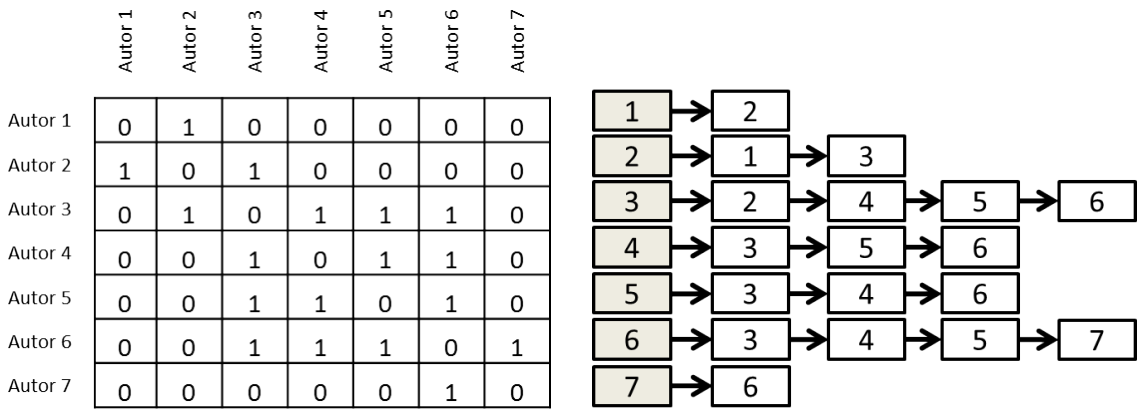
da aresta entre os respectivos vértices. Também podem possuir valores reais ou inteiros para representar o peso da aresta. Como consequência de um grafo não direcionado, no qual os relacionamentos são simétricos, a matriz é simétrica, isto é, a matriz triangular superior é igual a matriz triangular inferior. Seja um grafo $G(V, E)$, pode-se descrever uma matriz de adjacência M como:

$$m_{ij} = \begin{cases} w_{ij}, & \text{se} \quad (e_1, e_2) \in E, \\ 0, & \text{se} \quad \text{caso contrário.} \end{cases} \tag{1}$$

- Lista de adjacências (Figura 3(b)): as posições da lista são indexadas pelos vértices do grafo e cada uma dessas posições possui uma lista associada contendo os vértices adjacentes.

A utilização de matriz de adjacência pode ser não aplicável computacionalmente para grafos com uma quantidade muito grande de vértices, pois exige muito espaço de armazenamento. Adicionalmente, além de possuir muitos vértices, o grafo pode ser esparso, o que acarretaria em um grande desperdício de recurso computacional na representação matricial (CORMEN; LEISERSON; STEIN, 2012).

Figura 2 – Utilização da matriz de adjacência e a lista de adjacência para representar o grafo da Figura 1.



(a) Matriz de Adjacência (b) Lista de Adjacência

Fonte: William T. Maruyama, 2015.

2.2 Análise de Redes Sociais

Redes nas quais os elementos são pessoas ou grupos de pessoas e os relacionamentos são interações sociais entre os elementos são chamadas de redes sociais (NEWMAN, 2010;

ZAFARANI; ABBASI; LIU, 2014). É atribuído o surgimento deste termo ao antropologista John Arundel Barnes (WASSERMAN; FAUST, 1994). Apesar do termo "Redes Sociais" ter se popularizado atualmente devido aos serviços *online* disponíveis, entre os pesquisadores que estudam redes e os sociólogos há uma longa história de estudo nessa área. Sociólogos, por exemplo, desenvolveram uma linguagem própria para trabalhar com redes ao se referirem às pessoas como atores e aos relacionamentos entre as pessoas como laços (NEWMAN, 2010). Além disso, pode-se criar diferentes redes sociais agrupadas de acordo com o tipo de relacionamento criado entre as entidades (PRELL, 2011).

O campo de pesquisa que estuda as redes sociais é conhecido como Análise de Redes Sociais (ARS ou do inglês *Social Network Analysis* - SNA). Essa área procura estudar os relacionamento entre as entidades, ou seja, analisar os padrões ou regularidades e os resultados dessas relações (WASSERMAN; FAUST, 1994). ARS é uma área interdisciplinar que envolve esforços de diferentes áreas como: Antropologia, Sociologia, Psicologia, Matemática, Estatística e Ciência da Computação. Alguns conceitos na ARS surgiram de forma independente nos diferentes campos de pesquisa e por isto possuem termos diferentes de acordo com a área. Como exemplo, o termo "ator" vem da Sociologia, enquanto "nó" ou "vértice" vêm da Teoria dos Grafos (WASSERMAN; FAUST, 1994; PRELL, 2011).

Existem diferentes métricas que são utilizadas para a análise de uma rede, por exemplo, aquelas referentes à topologia da rede. A seguir são apresentadas algumas métricas típicas da análise de uma rede social e que podem ser encontradas em Zafarani, Abbasi e Liu (2014).

- Centralidade: métrica que procura determinar a importância de um nó na rede. Por exemplo, quando uma pessoa é influente em uma rede social, espera-se que sua centralidade possua um valor alto. Para essa finalidade, existem algumas medidas de centralidade que exploram características diferentes da rede. As métricas mais comuns são:
  - Centralidade de grau (*Degree Centrality*): é baseada na ideia de que pessoas com muitas conexões são mais centrais (ou importantes/influentes) do que as pessoas com menos conexões. Essa métrica é calculada como a quantidade de arestas incidentes a determinado vértice. Para um dígrafo temos a seguinte equação:

$$\text{grau}(v_i) = |\Gamma(v_i)| \tag{2}$$

Em que $\Gamma(v_i)$ representa o conjunto de vértices vizinhos ou adjacentes ao vértice v_i (dois vértices v_1 e v_2 são adjancetes se há uma aresta $e = (v_1, v_2)$ no grafo) e a quantidade de vizinhos ao vértice v_i é denotada por $|\Gamma(v_i)|$.

– Centralidade por intermediação (*Betweeness Centrality*): calcula a frequência com que um vértice é encontrado em um caminho geodésico[1] entre outros dois vértices. Pode ser descrita pela seguinte equação:

$$\text{intermediação}(v_i) = \sum_{s \neq t \neq v_i} \frac{\sigma_{st}(v_i)}{\sigma_{st}} \tag{3}$$

Sendo σ_{st} o número de caminhos geodésicos do vértice s até o vértice t, enquanto $\sigma_{st}(v_i)$ é o número de caminhos geodésicos do vértice s até o vértice t que passam pelo vértice v_i.

– Centralidade por proximidade (*Closeness Centrality*): é baseada na ideia de que os vértices mais centrais são aqueles mais próximos de todos os outros vértices da rede. É definida como:

$$\text{proximidade}(v_i) = \frac{1}{\bar{I}_{v_i}} \tag{4}$$

Sendo $\bar{I}_{(v_i)} = \sum_{v_i \neq v_j} dist(v_i, v_j)$ a média dos comprimentos dos caminhos geodésicos do vértice v_i para os outros vértices. Logo, quanto menor for o comprimento médio, maior será a centralidade do vértice.

- Transitividade: um tipo de análise realizada em uma rede é a formação de arestas (*links*) no grafo. Há duas medidas de agrupamento (*clustering*) tipicamente utilizadas para medir este comportamento:

  – Coeficiente local de agrupamento: procura estimar a frequência em que os vizinhos de um vértice estão conectados entre si. Sua representação matemática é:

$$\text{CL}(v_i) = \frac{\text{Número de pares de vizinhos de } v_i \text{ que estão conectados}}{\text{Número de pares de vizinhos de } v_i} \tag{5}$$

  – Coeficiente global de agrupamento: é baseado na contagem de triângulos (isto é, três vértices todos ligados entre si) presentes em toda a rede. Para o cálculo, pode-se fazer a contagem de caminhos de comprimento 2 e verificar se uma

[1] O caminho mais curto entre um par de vértices é chamado de geodésico, portanto, a distância geodésica é o comprimento ou tamanho desse caminho. O tamanho ou comprimento do caminho é dado pela quantidade de arestas (ligações) intermediárias entre o par de vértices que se deseja calcular a distância.

terceira aresta fecha do caminho (formando o triângulo). A seguinte equação descreve a métrica:

$$\text{CG} = \frac{\text{Número de triângulos} \times 3}{\text{Quantidade de conjuntos de três vértices conectados}} \quad (6)$$

2.3 Predição de Relacionamentos

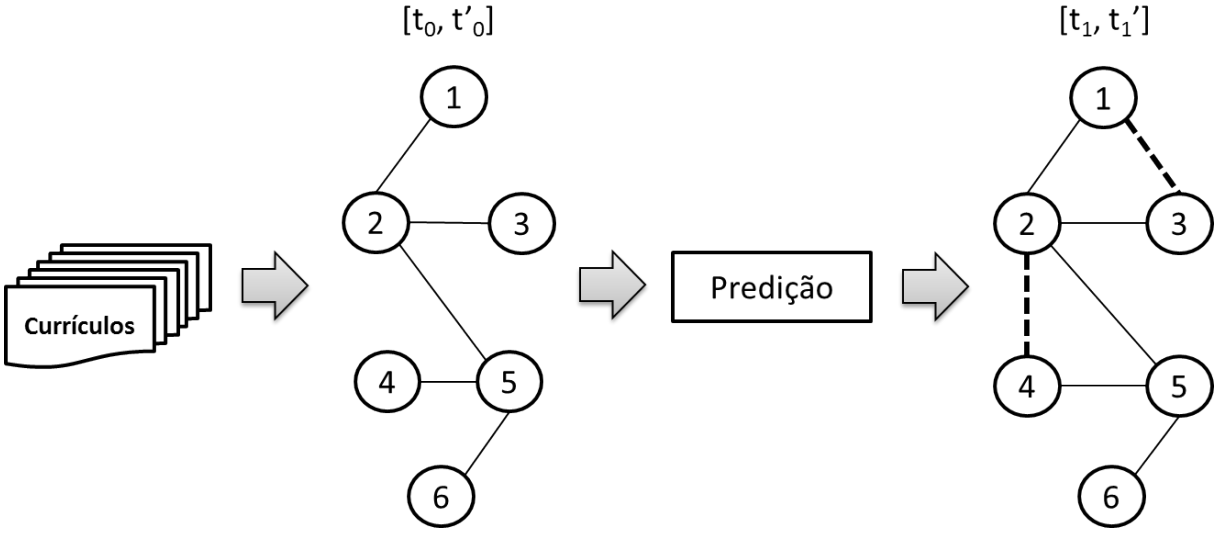
Conforme apresentado, os relacionamentos estabelecidos em uma Rede Social também podem ser chamados, genericamente, de *links*. A predição de relacionamentos é um tópico de estudo que tem como objetivo prever relacionamentos futuros que podem ser formados entre as entidades da rede ou servir de base para sugestão de novos relacionamentos.

Pode-se encontrar na literatura duas abordagens principais de predição de *links*. A primeira tenta predizer *links* faltantes, a fim de completar a estrutura da rede na qual há conexões omissas. A segunda abordagem é conhecida como predição temporal, pois se tenta predizer futuros relacionamentos que podem ou não existir no momento em que a rede é analisada, ou seja, tenta-se predizer a evolução da rede no futuro. Estudos de predição temporal de *links* podem ser definidos com a questão: dado um intervalo de tempo $[t_0, t'_0]$ da estrutura de uma rede, como prever as ligações da rede com precisão no futuro $[t_1,t'_1]$, sendo $t'_0 < t_1$? A Figura 3 ilustra a ideia geral da predição dos relacionamentos de coautorias, abordado no presente trabalho.

Para realizar a predição, utilizam-se métodos que medem a proximidade ou similaridade entre as entidades (nós) da rede. Esses métodos fornecem medidas que podem ser utilizadas por si só para predizer, mas podem ser adotadas como atributo ou característica a serem utilizadas por um sistema de mineração de dados (LU et al., 2010; SA; PRUDENCIO, 2011; SOARES; PRUDENCIO, 2012).

A predição de relacionamentos é um problema complexo, sendo que a identificação dos melhores conjuntos de atributos relevantes, dentre as várias combinações possíveis, é de suma importância para melhoria de precisão dos modelos preditivos (HASAN et al., 2006; HASAN; ZAKI, 2011). Além disso, tratar de conjuntos de dados tipicamente muito desbalanceados (dado um par arbitrário de pessoas há uma probabilidade grande de que elas não irão se relacionar) e o envolvimento de diversas técnicas estatísticas ou de inteligência artificial, em que cada técnica poderá apresentar melhores resultados de acordo

Figura 3 – Rede social acadêmica formada por relacionamentos de coautorias extraídas de informações dos currículos dos pesquisadores. Os relacionamentos preditos com linha tracejada são novos/inéditos e os com linha contínua são reincidentes.



Fonte: William T. Maruyama, 2015.

com o domínio no qual ela foi aplicada, também influenciam na complexidade da predição de relacionamentos.

Como o objeto de estudo da ARS são as redes sociais, é natural que algumas métricas tenham como base fundamentos da Sociologia, ao levar em conta o comportamento humano para o entendimento das interações dos indivíduos. Yin et al. (2010) apresentam alguns aspectos sociais para encontrar *links* relevantes:

1. Homofilia: quanto mais interesses em comum as pessoas possuem, como características ou preferências, maior as chances de se relacionarem.
2. Raridade: há mais chances de relacionamento entre pessoas com características ou preferências raras em comum. Pois características difíceis de encontrar tendem a se destacar em relação às outras características.
3. Influência Social: uma característica compartilhada com muitos amigos de uma determinada pessoa pode ser útil para encontrar potenciais relacionamentos.
4. Amizades em Comum (ou vizinhos em comum): quanto mais amigos em comum duas pessoas possuem, maiores são as chances de se relacionarem.
5. Proximidade Social: pessoas localizadas próximas em um grafo social possuem um relacionamento em potencial.
6. Conexão Preferencial: pessoas populares tendem a atrair mais pessoas quando comparadas às pessoas com poucos relacionamentos.

2.3.1 Conjunto de características

Existem variadas métricas para a predição de relacionamentos, dentre elas estão as métricas baseadas em similaridade (ou proximidade). As técnicas que utilizam essas métricas calculam para cada par de vértices da rede, uma pontuação (*score*), que representa sua similaridade. Os altos valores de pontuação indicam alta probabilidade de existência do *link* (LIU et al., 2005; HASAN; ZAKI, 2011).

Uma métrica muito adotada é a similaridade estrutural em que são extraídas informações (padrões) topológicas do grafo (local ou global). Em geral, essas métricas são adaptadas de técnicas usadas na Teoria de Grafos e ARS (LIBEN-NOWELL; KLEINBERG, 2003; LÜ; ZHOU, 2010).

Lü e Zhou (2010) e Hasan e Zaki (2011) realizaram levantamentos de diversos atributos baseados na topologia do grafo, para predição de *links*. Nesses levantamentos, os autores dividiram os atributos em dois conjuntos. A seguir, são apresentados os atributos tipicamente utilizados, considerando um grafo não ponderado e não direcionado.

O primeiro conjunto de atributos foi denominado como conjunto de métricas baseadas em vizinhança (HASAN; ZAKI, 2011) ou métricas de índice local (LÜ; ZHOU, 2010). Elas são métricas calculadas com base na informação local de um nó, isto é, utiliza-se a informação da estrutura dos nós vizinhos:

- Vizinhos em Comum (*Common Neighbors*): quantidade de vizinhos em comum entre um par de nós. À medida que a quantidade de vizinhos em comum cresce, a chance dos dois nós terem um *link* entre si aumenta.

$$\mathrm{CN}(x, y) = |\Gamma(x) \cap \Gamma(y)| \tag{7}$$

  Sendo $\Gamma(x)$ o conjunto de vizinhos do vértice x e $|\Gamma(x)|$ a quantidade de vizinhos do vértice x, isto é, o grau do vértice.

- Jaccard *Coefficient* (JACCARD, 1901): é uma métrica de similaridade que normaliza a quantidade de vizinhos em comum.

$$\mathrm{JC}(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \tag{8}$$

- Adamic-Adar (ADAMIC; ADAR, 2001): é uma métrica de similaridade que pondera a vizinhança em comum entre os vértices. Para isso, atribui maior peso aos vizinhos em comum com menor quantidade de arestas.

$$\mathrm{AA}(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log|\Gamma(z)|} \tag{9}$$

- Conexão Preferencial (*Preferential Attachment*): Newman (2001) e Barabási et al. (2002) propõem que as chances de um vértice estar conectado a outro vértice em uma rede é proporcional ao produto do número de vizinhos que cada um possui. A ideia é que novos relacionamentos têm mais chances de correr com pessoas que têm muitos relacionamentos (conhecido como o rico fica mais rico)(HASAN; ZAKI, 2011).

$$\mathrm{PA}(x,y) = |\Gamma(x)|.|\Gamma(y)| \tag{10}$$

- Salton ou Similaridade Cosseno (SALTON; MCGILL, 1986): é uma medida da similaridade entre dois vetores que mede o cosseno do ângulo entre os mesmos.

$$\mathrm{SA}(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{|\Gamma(x)|.|\Gamma(y)|}} \tag{11}$$

- Sørensen (SØRENSON, 1948): é usada para medir a similaridade entre duas amostras, frequentemente empregado em análises de comunidades ecológicas.

$$\mathrm{SO}(x,y) = \frac{2|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x)| + |\Gamma(y)|} \tag{12}$$

- *Hub Promoted Index* (GIRVAN; NEWMAN, 2002): foi proposta para medir a sobreposição topológica de pares de substratos de redes metabólicas.

$$\mathrm{HPI}(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{\min\{|\Gamma(x)|, |\Gamma(y)|\}} \tag{13}$$

- *Hub Depressed Index* (ZHOU; L; ZHANG, 2009): oposta a métrica anterior.

$$\mathrm{HDI}(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{\max\{|\Gamma(x)|, |\Gamma(y)|\}} \tag{14}$$

- Leicht-Holme-Newman (LEICHT; HOLME; NEWMAN, 2006): medida de similaridade proposta com base no conceito de que dois vértices são semelhantes se os seus vizinhos na rede também são semelhantes.

$$\mathrm{LHN}(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x)|.|\Gamma(y)|} \tag{15}$$

- *Resource Allocation* (OU et al., 2007): métrica que, dado um par de vértices não conectados diretamente, tenta mensurar a transmissão de recursos entre estes vértices. Para um vértice x enviar um recurso para o vértice y, tem-se a utilização dos vértices vizinhos para transmissão. No caso mais simples, assume-se que cada transmissor possui uma unidade de recurso e a distribuição é feita igualmente para todos os vizinhos. Portanto a similaridade é dada pela quantidade de recursos que y recebe de x.

$$\mathrm{RA}(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{|\Gamma(z)|} \tag{16}$$

O outro conjunto de atributos foi chamado de conjunto de métricas baseadas no caminho (HASAN; ZAKI, 2011). Lü e Zhou (2010) chamaram essa categoria de conjunto métrica de índice global. Essas métricas utilizam informação global da rede, ou seja, o cálculo considera os caminhos possíveis entre um par de nós.

- Menor distância (*Shortest Path*): distância mais curta entre os pares de nós em uma rede. Métrica baseada na hipótese de que a distância entre os nós e a probabilidade de estarem conectados são inversamente proporcionais. Então, Liben-Nowell e Kleinberg (2003) incluem um fator negativo no menor caminho.

$$\mathrm{SP}(x,y) = -dist(x,y) \tag{17}$$

Sendo a função *dist* a menor distância entre os vértices x e y.

- Katz (KATZ, 1953): variação da menor distância. Realiza a soma de todos os caminhos possíveis entre os vértices da rede e pondera com mais peso os caminhos mais curtos.

$$\mathrm{Katz}(x,y) = \sum_{l=1}^{\infty} \beta^l . |paths_{x,y}^{(l)}| \tag{18}$$

Sendo $paths_{x,y}^{(l)}$ o conjunto de todos os caminhos de tamanho l, com x como vértice de origem e y o vértice de destino. $\beta \in (0,1)$ é a constante que pondera de acordo com a relevância do tamanho do caminho. Alternativamente, pode-se calcular a matriz de *scores* dessa métrica a partir da matriz de adjacência.

$$\mathrm{Katz} = (I - \beta M)^{-1} - I \tag{19}$$

Em que M é matriz de adjacência do grafo e I a matriz identidade.

- *Hitting time*: são realizados percursos aleatórios em um grafo. O *Hitting time* entre dois vértices x e y do grafo é dado pelo número esperado de passos necessários de um percurso aleatório para alcançar y a partir de x.
- *Rooted Pagerank*: baseada na adoção da métrica *Pagerank* (LANGVILLE; MEYER, 2009). Por meio de um percurso aleatório entre dois vértices no qual, a cada passo, é definida a continuidade ou o reinício do processo a partir de uma probabilidade. Uma forma de calcular é a utilização da forma matricial, na qual será obtida a matriz de *scores*.

$$\text{RPR} = (1 - \alpha) \times (I - (\alpha.N))^{-1} \tag{20}$$

Sendo N igual a $D^{-1}M$, M a matriz de adjacência do grafo e D uma matriz diagonal, definida como $D_{i,i} = \sum_j M_{i,j}$. A probabilidade do processo ser reiniciado é dado por α e o processo para selecionar de forma aleatória um vizinho a partir do vértice atual é dado pela probabilidade $(1 - \alpha)$.

Segue um exemplo do cálculo de algumas métricas para o par de vértices/autores 3 e 4 da Figura 1. Tem-se que $\Gamma(3) = \{2, 4, 5, 6\}$ e $\Gamma(4) = \{3, 5, 6\}$, logo $CN(3, 4) = 2$, $JC(3, 4) = \frac{2}{7}$ e $PA(3, 4) = 27$.

As métricas descritas possuem a vantagem de serem genéricas, isto é, podem ser aplicadas em diversos contextos, por exemplo, nas redes de amizade e de coautoria. Elas podem ser consideradas genéricas, pois são baseadas na topologia da rede sem considerar nenhuma outra característica específica dos vértices ou das arestas.

Diferentemente, as métricas baseadas nas características dos vértices variam de acordo com o contexto da rede em que as informações são extraídas considerando a semântica ou conteúdo associado ao vértice. Essas características possuem a vantagem de considerar fatores e padrões que são inerentes aos elementos da rede. Contudo, é necessário um bom conhecimento do contexto da aplicação para identificação das características (HASAN; ZAKI, 2011) dos vértices. Além disso, em alguns casos, a exploração das informações dos elementos da rede não é permitida devido a segurança ou privacidade desses elementos. Exemplos de características de contexto serão apresentadas na Tabela 5.

A Tabela 1 apresenta um resumo da principal vantagem e desvantagem dos tipos de atributos utilizados.

Para calcular a similaridade entre dois vértices da rede e considerar a informação particular de cada elemento, os autores Hasan et al. (2006) utilizaram a métrica baseada

Tabela 1 – Vantagens e desvantagens entre atributos estruturais e de domínio/contexto.

	Vantagem	Desvantagem
Domínio/Contexto	Considerar fatores e padrões que são inerentes aos elementos da rede.	Variam de acordo com o contexto da rede. Informação restrita.
Estrutural	Consideradas genéricas.	Para as grandes redes sociais, algumas dessas características podem ser computacionalmente custosas.

Fonte: William T. Maruyama, 2015.

em agregação de características. Essa métrica utiliza uma função que irá produzir um valor (*score*) de similaridade com as informações dos vértices. As funções de agregação podem ser, por exemplo: soma, média, máximo, mínimo, etc. No caso de Hasan et al. (2006), a função soma foi utilizada como função agregativa em duas redes de coautoria para os pares de autores, aplicando a função para criar as características como: Soma de publicações, Soma de vizinhos e Soma da contagem das palavras chaves.

2.3.2 Estratégias

Com os valores obtidos das métricas calculadas, é necessária uma estratégia de avaliação para decidir quais pares de nós possuem um relacionamento. Para tanto, existem duas estratégias que podem ser utilizadas: a supervisionada e a não supervisionada.

2.3.2.1 Método Supervisionado

A predição de *links* utilizando a estratégia supervisionada é tipicamente tratada como um problema de classificação binária, em que os *links* entre os pares de nós possuem duas condições (rótulos ou classificações) possíveis: presença (1 ou classe positiva) e ausência (0 ou classe negativa) (HASAN et al., 2006; HASAN; ZAKI, 2011).

Os *links* da rede são divididos de modo que seja obtido um conjunto de dados para treinamento e um conjunto de dados para teste. A partir das redes são extraídas ou calculadas as características (como métricas de similaridade e de contexto) e é construído o vetor de características $\vec{v} = \{x_1, x_2, x_3, \ldots, x_n\}$, em que cada vetor de características (instância) é rotulado, isto é, cada instância possui uma classe.

O conjunto de treinamento, por sua vez, é utilizado para treinar um classificador, como Rede Neural, Árvore de Decisão, Máquina de Vetores de Suporte (do inglês, *Support Vector Machine* - SVM), etc. Com o classificador treinado, pode-se obter um modelo preditivo, no qual novas instâncias não rotuladas são submetidas para serem classificadas. O conjunto de teste é utilizado para avaliar os modelos treinados.

No presente trabalho, foi utilizado está estratégia para a predição de relacionamentos de coautorias.

2.3.2.2 Método Não Supervisionado

Nesse método, são utilizados diretamente os valores calculados pelas métricas para a predição de *links*, não tendo uma etapa de treinamento de um modelo preditivo (LÜ; ZHOU, 2010).

Os *links* observados da rede são divididos, um conjunto para a análise e outro para os testes. O conjunto de dados de análise é utilizado para extração das características (cálculo dos *scores*) pertencentes aos pares de vértices. Enquanto que o conjunto de dados de teste é utilizado para avaliação do desempenho da predição. Para divisão do conjunto de dados duas estratégias de amostragem podem ser utilizadas, como a validação cruzada ou a sub-amostragem aleatória (LÜ; ZHOU, 2010; LIN; YUN; ZHU, 2012).

Inicialmente os *scores* calculados das métricas no conjunto em análise são ordenados de maneira decrescente, de forma que os maiores *scores* dos pares de vértice ficam no topo da lista. Em seguida, os pares de vértices do conjunto de teste são avaliados segundo a definição das ligações.

Existem duas abordagens para considerar quais pares estarão conectados:

- Uma abordagem é definir um limiar θ, sendo que pares de nós que tiverem os *scores* acima do limiar serão considerado como conectados;
- Outra abordagem é definir a quantidade de *links* da lista dos *scores* dos pares de nós que serão considerados conectados.

Para encontrar o melhor resultado, nas duas abordagens, são testados diversos valores para o limiar ou a quantidade de *links*. Lü e Zhou (2010) e Lin, Yun e Zhu (2012) indicam as métricas Área sob a curva e Precisão na avaliação do desempenho de métodos de predição na estratégia não supervisionada.

2.4 Mineração de Dados

O termo Mineração de Dados (*Data Mining*) caracteriza o processo de análise de uma grande quantidade de dados, no qual pode ser encontrado conhecimento relevante e padrões (HAN; KAMBER, 2012). A Mineração de Dados possui uma natureza interdisciplinar, pois possui relações com, por exemplo, estatística, aprendizado de máquina, reconhecimento de padrões e sistemas de banco de dados. Nesse contexto, alguns autores consideram o termo Descoberta de Conhecimento a Partir de Dados (do inglês *Knowledge Discovery from Data* - KDD) como um sinônimo de Mineração de Dados, enquanto outros tratam a mineração de dados como um passo no processo de descoberta de conhecimento (HAN; KAMBER, 2012).

Identificar padrões de um conjunto de objetos possibilita a capacidade de discriminar (de conseguir classificar) um objeto de entrada a partir de suas características mais significativas dentre todos os possíveis. Portanto, as características servem para descrever particularidades de um objeto de estudo. No contexto de uma rede social, são os relacionamentos estabelecidos que geralmente apresentam padrões em um conjunto de dados (GETOOR; DIEHL, 2005). Deste modo, o termo *Link Mining* explicita a utilização de técnicas de Mineração de Dados sobre esses *links* (GETOOR; DIEHL, 2005), com o intuito de descobrir esses padrões.

A análise de padrões das redes sociais é uma das bases para a predição de *links*, que tem como objetivo prever relacionamentos futuros que podem ser formados entre as entidades da rede ou servir de base para sugestão de novos relacionamentos. Neste tópico será apresentada uma breve introdução sobre as etapas de pré-processamento e processamento de dados. A etapa de pré-processsamento terá foco em atividades de normalização, seleção de atributos e extração de características. Na etapa de processamento serão apresentadas algumas técnicas de classificação que podem ser utilizadas para a predição.

2.4.1 Normalização

Em alguns casos, os dados utilizados apresentam unidades diferentes, o que pode influenciar os resultados das técnicas de classificação aplicadas. A normalização dos dados é, portanto, usada com o intuito de evitar as tendências de diferentes escalas de valores

das características. Para tal, utiliza-se uma função que realiza o mapeamento dos valores de atributos para um novo conjunto de valores. Uma função de normalização que trata a amplitude dos valores é:

$$\text{valor\_normalizado} = \frac{\text{valor} - \min}{\max - \min} \cdot (\text{novo\_max} - \text{novo\_min}) + \text{novo\_min} \tag{21}$$

Denominada normalização Min-Max (HAN; KAMBER, 2012), na qual "min" e "max" correspondem, respectivamente, ao valor mínimo e máximo encontrado no intervalo de valores de um atributo. Enquanto que "novo_min" e "novo_max" são os novos valores mínimo e máximo do intervalo. Isto é, os valores normalizados estarão em um intervalo [n_min, n_max]. Esta função mantém a mesma relação entre as instâncias, conforme o conjunto de atributos originais, contudo a presença de *outliers* pode influenciar o resultado.

Outra maneira de normalizar os dados é utilizar uma normalização distribucional. Essa normalização pode ser interessante, por exemplo, para obter a simetria dos atributos a partir de sua média (μ) e seu desvio padrão (σ). Pode-se utilizar a seguinte função:

$$\text{valor\_normalizado} = \frac{\text{valor} - \mu}{\sigma} \tag{22}$$

O atributo normalizado (v') terá média zero e desvio padrão igual a 1. Esta função é útil quando não se conhece o máximo e o mínimo do atributo real, ou quando há valores discrepantes que possam interferir no intervalo da normalização (HAN; KAMBER, 2012).

2.4.2 Seleção de atributos

O objetivo desta atividade é identificar as características mais relevantes no conjunto de dados, podendo eliminar características que pouco contribuem (ou são redundantes) na classificação. Essa atividade pode evitar um problema que pode ocorrer ao se analisar muitas características – isto é, a alta dimensionalidade –, denominado como Maldição da Dimensionalidade (BISHOP, 2006). Esse problema implica que, para um tamanho de amostra, existe um número máximo de características a partir do qual o desempenho do classificador irá degradar ao invés de melhorar. Ele também é conhecido por curva em U, figura formada a partir do desempenho do classificador plotado em um gráfico. Portanto,

essa redução pode melhorar a eficácia e eficiência dos classificadores, reduzir o tamanho da amostra e facilitar a visualização dos dados.

Algoritmos realizam de diferentes maneiras a seleção das características, os quais podem ser divididos segundo sua relação com o classificador. Em Dash e Liu (1997), os métodos de seleção de características são categorizadas em três grandes grupos:

- Filtros: métodos independentes do classificador, geralmente utilizados antes do processo de classificação, de forma a selecionar o subconjunto de atributos;
- Camadas (*wrappers*): métodos dependentes do classificador, em que é utilizada uma camada sobre o mesmo ("caixa preta"). O algoritmo de seleção faz chamadas ao classificador para realizar a avaliação, com o intuito de encontrar um subconjunto de atributos que melhor se adequa ao algoritmo;
- Embutidos: é uma abordagem realizada durante o processo de treinamento do classificador, já utilizando algum critério interno de alguns algoritmos de seleção. Exemplo: Árvore de Decisão.

No presente trabalho, para montar os diferentes subconjuntos de dados para os experimentos foram utilizados algoritmos do tipo filtro.

Testar todos os subconjuntos possíveis e escolher a melhor combinação é um problema exponencial, pois teremos 2^n (sendo n a quantidade de características) subconjuntos possíveis para serem testados. Para percorrer o espaço de busca e gerar os subconjuntos existem algumas estratégias (DASH; LIU, 1997) que podem ser agrupados da seguinte forma:

- Completo: este procedimento realiza uma pesquisa completa pelo subconjunto ideal de acordo com uma função de avaliação. O método é exaustivo, realizando o teste em todas as possibilidades. No entanto, diferentes funções heurísticas são usadas para reduzir a procura, sem comprometer as chances de encontrar o subconjunto ideal. A otimização, de acordo com a função de avaliação, é garantida pelo procedimento de *backtracking*, o qual pode ser feito ao utilizar técnicas como: *Branch and Bound* e Busca Iniciando pelo Melhor (do inglês, *Best-first search*);
- Heurístico: é um processo basicamente incremental. A cada iteração, uma nova característica é adicionada ao subconjunto ou inicialmente é utilizado todo o conjunto, e a cada iteração uma característica é removida. Um exemplo é o algoritmo *Relief*;

- Randômico: é um procedimento de geração aleatória. A busca pelo subconjunto ideal neste método depende dos recursos disponíveis e nele é definido um número máximo de iterações possíveis.

Algumas métricas podem ser utilizadas para encontrar o melhor subconjunto. Para isso, é utilizada uma função de avaliação chamada de função critério, cujo objetivo é buscar o melhor elemento dentro do conjunto de acordo com uma métrica estipulada. Logo, ao escolher um subconjunto esse deverá ser ótimo em relação à função utilizada. Podem ser utilizadas as seguintes medidas:

- Medidas de distância entre classes distintas: podem ser utilizados os cálculos de distância mínima, distância máxima, distância média, distância média de todos os pares de classes distintas e distância entre os centros de massa.
- Medidas de informação:
  - Informação mútua: avalia a informação mútua entre subconjuntos e a classe, e compara qual deles tem maior quantidade de informação, pois isso indica que é possível predizer a classe de maneira mais precisa com aquele conjunto de características;
  - Entropia: ao medir a entropia de um subconjunto e a classe, calcula-se o nível de incerteza de saber a classe ao ver aquele conjunto de características. Então, quanto menor a entropia de um subconjunto de características em relação à classe, melhor ele será.
- Medidas de dependência:
  - Correlação: a escolha pode ser realizada com base na maior correlação entre a característica e a classe. Ou então se duas características tiverem uma correlação alta uma delas pode ser descartada;
  - Medidas de consistência: uma característica é consistente se todas as instâncias com um determinado valor da característica possuem a mesma classificação;
  - Taxa de erro do classificador: do tipo *wrapper*, a escolha será aquele subconjunto que apresenta menos erro.

2.4.3 Extração de atributos - redução de dimensionalidade

Assim como na seleção de características, a extração pode diminuir a dimensão do espaço de características. No entanto, diferentemente da seleção, a extração realiza a tranformação ou a combinação dos atributos do espaço original para obter um novo espaço de atributos e com possibilidade de diminuição do mesmo. Uma técnica bem conhecida é a análise de componentes principais (do inglês *Principal Component Analysis* - PCA).

Com o PCA, as características originais são transformadas de maneira que novas características sejam criadas (com mudança da base do espaço vetorial) e essas tenham nos primeiros componentes principais (PCs) as maiores variâncias. A redução é obtida por meio do estabelecimento de novas características ortogonais entre si, denominadas componentes principais. Organizadas em ordem decrescente de importância, as PCs são combinações lineares das características originais.

2.4.4 Classificação

Dentro da mineração de dados há a atividade de classificação, na qual se busca um modelo ou função que descreve e distingue as classes (ou conceitos). Os modelos são derivados da análise do conjunto de dados de treinamento. Esse processo é conhecido como aprendizado supervisionado, pois os classificadores necessitam conhecer as classes de cada instância da amostra para o processo de aprendizagem do modelo. O modelo é usado para prever a classe das instâncias desconhecidas.

Existem diversos métodos propostos na literatura, que podem ser aplicados na classificação. Neste projeto foram utilizados apenas algoritmos cuja implementação está disponível no arcabouço Weka[2]. A seguir serão sumarizados os algoritmos utilizados, agrupados conforme sua base de aprendizado.

2.4.4.1 Redes Bayesianas

As Redes Bayesianas são algoritmos de Aprendizado de Máquina capazes de fornecer predições associadas aos valores de probabilidades.

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

- *Naive Bayes Simple*: é a implementação simplista do *Naive Bayes*, que é um classificador probabilístico. O classificador é denominado ingênuo, pois assume que os atributos são independentes. Esse modelo aplica o Teorema de Bayes para estimar cada valor de classe a partir dos valores dos atributos. Os atributos numéricos são modelados por uma distribuição normal (DUDA; HART, 1973).
- *Naive Bayes* (NB): é a implementação do algoritmo *Naive Bayes* usando estimador de classes. Os valores de precisão dos estimadores numéricos são escolhidos com base na análise dos dados de treinamento. Mais detalhes podem ser encontrados em John e Langley (1995);
- *Naive Bayes Updateable*: é a versão atualizável do algoritmo *Naive Bayes*, na qual o estimador de classes é atualizado durante o treinamento. Este classificador utiliza uma precisão padrão de 0,1 para os atributos numéricos quando *buildClassifier* for chamado com zero instâncias de treinamento (JOHN; LANGLEY, 1995);
- *DMNBtext*: Su et al. (2008) propuseram um algoritmo de classificação de texto, chamado *Discriminative Multinomial Naive Bayes* (DMNB), que leva em conta tanto a probabilidade e os objetivos de classificação durante a contagem de frequência;
- *Bayesian Logistic Regression*: implementa Regressão Logística Bayesiana para Gaussiana e *Laplace Priors* (GENKIN; LEWIS; MADIGAN, August 2007). O *Laplace Priors* é utilizado para evitar *overfitting* (superajustamento ou superespecialização) e produz modelos preditivos em dados de texto esparsos.
- *Bayes Net*: Redes Bayesianas pertencem à família de modelos probabilísticos em grafos, as quais codificam as relações probabilísticas entre as variáveis de interesse. Estes grafos são usados para representar o conhecimento sobre um domínio. Cada nó do grafo representa uma variável aleatória, enquanto as arestas entre os nós representam dependências probabilísticas entre as variáveis aleatórias correspondentes. Essas dependências condicionais no grafo são frequentemente estimadas por meio de métodos estatísticos (BOUCKAER, 2008).

2.4.4.2 Aprendizado de funções

São algoritmos que buscam modelar uma função que se aproxime do mapeamento dos dados de entrada. Para isso, os coeficientes das funções são aprendidos na fase de aprendizado.

- *Voted Perceptron*: Freund e Schapire (1999) propuseram um algoritmo baseado no algoritmo Perceptron. É um algoritmo simples para classificação linear, cuja ideia é encontrar a maior margem que separe dois conjuntos de dados. Além disso, esse classificador utiliza funções *kernel* para adicionar dimensões, pois alguns problemas não são linearmente separáveis, na dimensão do conjunto de características de entrada;
- *Simple Logistic*: o algoritmo constrói modelos de regressão logística lineares. Para a montagem dos modelos logísticos, é utilizado o algoritmo *LogitBoost* com funções de regressão simples como base para o aprendizado (LANDWEHR; HALL; FRANK, 2005; SUMNER; FRANK; HALL, 2005);
- Otimização Mínima Sequencial (do inglês, *Sequential Minimal Optimization* - SMO): Proposto por Platt (1999), este algoritmo treina um classificador de vetores de suporte. O SMO se propõe a resolver de forma eficiente o problema de programação quadrática relacionado ao processo de treinamento do classificador SVM. Diferentemente do SMO, o algoritmo SVM possui um treinamento demorado para grandes conjuntos de dados e o algoritmo de treinamento é mais complexo o que pode tornar a implementação mais difícil. Além disso, SMO possui um algoritmo conceitualmente mais simples, fácil de implementar e geralmente mais rápido do que a implementação tradicional do algoritmo de treinamento do SVM;
- *RBF Network*: a rede RBF (do inglês *Radial Basis Function*, função de base radial) é uma classe de modelos de Redes Neurais. A função de ativação das unidades da camada oculta é definida por uma função de base radial, para calcular a distância entre o vetor de entrada e um vetor de protótipo. O algoritmo utiliza uma função Gaussiana normalizada. Além disso, utiliza o algoritmo de agrupamento *K-Means* no treinamento para determinar os protótipos;
- *Multilayer Perceptron* - (MLP): é uma rede Perceptron com pelo menos uma camada oculta (ou intermediária). É dita progressiva (*feedforward*), pois as saídas dos neurônios de uma camada se conectam unicamente às entradas dos neurônios da camada seguinte, sem a presença de laços de realimentação. Ademais, a rede possui uma alta conectividade e os neurônios da camada intermediária utilizam uma função de ativação não linear. O treinamento da rede utiliza o algoritmo denominado de retropropagação do erro;

- *Logistic*: é uma implementação alternativa para construção e uso de modelos de regressão logística multinominais com um *Ridge Estimator* com intuito de prevenir superajustamento (CESSIE; HOUWELINGEN, 1992). O algoritmo de regressão logística original foi modificado para lidar com os pesos nas instâncias.

2.4.4.3 Aprendizado Preguiçoso

A ideia chave do aprendizado baseado em instâncias é que a classe de uma instância de teste é, provavelmente, a mesma de exemplos com valores de atributos similares. O aprendizado baseado em instâncias também é conhecido aprendizado preguiçoso, já que o aprendizado consiste apenas em armazenar os exemplos de treinamento.

- *Locally Weighted Learning* (LWL): o algoritmo atribui pesos para as instâncias de treinamento. O peso expressa a influência da instância na predição. A ideia básica do LWL é que em vez de construir um modelo global para todo o espaço funcional, para cada instância de interesse um modelo local é criado com base em dados da vizinhança da instância observada. Em geral, os dados que estão na vizinhança próxima ao dado consultado atual recebe um peso maior do que dados que estão longe (ATKESON; MOORE; SCHAAL, 1997). Pode-se utilizar para classificação, por exemplo NB (FRANK; HALL; PFAHRINGER, 2003);
- *KStar*: proposto por Cleary e Trigg (1995), baseia-se na classe das instâncias de formação semelhante, conforme determinado por uma função de similaridade. Utiliza funções de distância baseadas na entropia e assume que os exemplos similares terão classes similares;
- *IB1*: Este algoritmo é baseado no vizinho mais próximo (AHA; KIBLER; ALBERT, 1991). Para isso, utiliza a distância euclidiana normalizada para encontrar a instância de treinamento mais próxima da instância de teste. A instância mais próxima é usada para classificar a instância de teste. Caso muitas instâncias de treinamento tenham a mesma distância o critério utilizado é da primeira encontrada.

2.4.4.4 Aprendizado Baseado em Regras

São algoritmos que determinam um conjunto de regras que formam um relacionamento entre os atributos e as classes. Uma regra é composta de duas partes: consequente e corpo (ou antecedente), no qual consequente é a classe predita e o corpo, uma conjunção de antecedentes em que cada um deles é uma condição que envolve um único atributo. Alguns algoritmos são:

- *ZeroR*: é um método de classificação simples que ignora os atributos e foca-se na classe. A classificação é dada pela classe majoritária. Não possui poder de previsibilidade, mas é útil para determinar uma referência de desempenho para outros métodos de classificação;
- *JRIP*: é um algoritmo de extração direta de regras que implementa a Poda Incremental Repetida para Produzir Redução de Erro (do inglês, *Repeated Incremental Pruning to Produce Error Reduction* - RIPPER), proposto por Cohen (1995);
- *Decision Table*: é um classificador simples que realiza a seleção de atributos que conduzem ao melhor resultado. Permite uma representação tabular das regras de decisão. Essa representação é formada por condições, ações e regras;
- *DTNB*: é um classificador híbrido que combina NB com a Tabela de Decisão (DT) que representa as probabilidades condicionais (HALL; FRANK, 2008). É usada uma busca de seleção para frente, em que a cada passo, o algoritmo avalia o ponto de divisão dos atributos, separando-os em dois subconjuntos disjuntos - um é modelado pelo NB e o outro é modelado pela Tabela de Decisão. Com a regra de Bayes, são combinadas as estimativas de probabilidade da classe dos dois modelos para realizar as estimativas globais de probabilidade da classe;
- *Conjunctive Rule*: é um classificador que implementa uma única regra, que pode prever rótulos numéricos ou categóricos. Ela é composta de antecedentes "AND" e de consequente ("valor da classe") para realizar a classificação. Se a regra não atender uma dada instância de teste, então a classe será prevista usando a distribuição de classe padrão. Na fase de aprendizado é selecionado um antecedente pelo cálculo do ganho de informação para cada antecedente e realizada a poda da regra gerada usando o REP (*Reduced Error Prunning*) ou a simples pré-poda baseada no número

de antecedentes. Para a classificação, a informação de um antecedente é a média ponderada das entropias de ambos os dados: abrangidos e não abrangidos pela regra.

2.4.4.5 Árvores de Decisão

A Árvore de Decisão (árvore de classificação ou apenas AD) é um modelo de classificação. O conhecimento adquirido, obtido por um processo de aprendizado supervisionado, pode ser construído em linguagem de alto nível a partir de uma representação simbólica. Logo, a AD facilita a interpretação pelas pessoas.

Uma característica das ADs é sua estrutura hierárquica, na qual a representação é uma árvore invertida. Portanto, se inicia a árvore pelo nó raiz e termina em nós folhas. Essa hierarquia é formada por um conjunto de elementos chamados de nós e suas relações de paternidade representadas por ramos (ligações).

Sua estrutura é definida no processo de indução da árvore de decisão, com divisões recursivas dos exemplos em subconjuntos menores na tentativa de separar cada classe. Logo, o algoritmo de aprendizado de uma AD tem como ponto central o critério utilizado para escolha do atributo que irá particionar o conjunto de exemplos a cada iteração. Após a construção da árvore, a classificação de um novo exemplo começa pela raiz da árvore, seguindo cada nó interno de decisão de acordo com o valor do atributo do novo exemplo até que uma folha seja alcançada.

- *Classification and Regression Trees* (CART): consiste de uma técnica não-paramétrica que induz tanto árvores de classificação quanto árvores de regressão, dependendo se o atributo é nominal (classificação) ou contínuo (regressão). O resultado é uma árvore binária, na qual cada nó testa um atributo gerando apenas dois ramos. A estratégia adotada para a realização da partição do conjunto de dados é por meio da medida de impureza. A ideia é dividir em subconjuntos de exemplos mais puros, para isso é utilizado o índice Gini. Além disso, utiliza a pós-poda por meio da redução do fator custo-complexidade. A técnica de poda utilizada produz árvores mais simples, precisas e com boa capacidade de generalização (BREIMAN et al., 1984).
- *Random Tree*: constrói cada nó da árvore escolhendo um entre K atributos selecionados aleatoriamente e testa cada um com o índice Gini. Além disso, não há

necessidade de poda e uma folha só é criada se nenhum dos atributos apresentar redução na quantidade de informação necessária a classificação dos exemplos.

- *Random Forest*: o conjunto de treinamento é divido aletoriamente em n subconjuntos diferentes, sendo que cada subconjunto é utilizado para construir uma árvore de decisão. Cada instância do conjunto de teste é classificada pelas árvores de decisão criadas, sendo que os rótulos são decididos por votação majoritária (BREIMAN, 2001);
- *REPTree*: constrói uma árvore de decisão que pode ser usada tanto para problemas de classificação quanto de regressão e utiliza o ganho de informação/variância. Ademais, usa a poda de redução de erro com ajuste retroativo (*backfitting*);
- *NBTree*: é um classificador híbrido entre uma árvore de decisão e um classificador NB. Em uma árvore construída com o *NBTree*, uma instância é classificada usando um NB na folha (KOHAVI, 1996);
- *LADTree*: é uma extensão do algoritmo ADTree. Esse algoritmo possibilita tratar problemas multiclasse utilizando a estratégia *LogitBoost*. Para tanto, o algoritmo divide o problema multiclasse em vários problemas de duas classes (HOLMES et al., 2001);
- *C4.5*: é uma implementação denominada como J48 no Weka e é a melhoria do algoritmo ID3. O algoritmo escolhe o atributo ponderando o ganho de informação esperado em relação ao nó pai. Uma das melhorias em relação ao ID3 é aplicação da técnica de pós-poda da árvore para combater o problema de superajustamento do conjunto de treinamento (QUINLAN, 1993);
- *J48graft*: utiliza o algoritmo J48 para criação da AD e aplica uma técnica de *grafiting*. A técnica adiciona novos nós na árvore treinada para redução do erro de classificação (WEBB, 1999);
- *Functional Tree* (FT): cria árvores de classificação, realizando a combinação de atributos (multivariante) nos nós e nas folhas. Usa-se regressão linear para criar combinações lineares dos atributos, durante o crescimento da árvore, e folhas durante o processo de poda. A implementação do FT no Weka pode usar função de regressão logística nos nós internos e/ou folhas (GAMA, 2004; LANDWEHR; HALL; FRANK, 2005).
- *Decision Stump*: cria uma AD com uma camada por atributo;
- *Best-First decision Tree* (BFTree) é parecido com o algoritmo C4.5. A principal diferença está na ordem da construção dos nós. Enquanto o C4.5 expande os nós em

ordem fixa da esquerda para direita, o BFTree procura construir começando pelo nó com a máxima redução de impureza para realizar a divisão. Essa estratégia é chamada de *best-first order* e pode evitar especialização da árvore durante o processo de construção (SHI, 2007).

2.4.4.6 Meta-Classificadores

Sistemas de aprendizagem meta-classificadores operam na saída de outros algoritmos de aprendizagem. A ideia básica é a combinação de vários modelos para formar um conjunto de classificadores para decidir sobre a classificação.

- *Vote*: nele é possível selecionar um conjunto de classificadores e combinar os resultados de probabilidade, a partir de uma determinada regra (majoritário, média, máximo, mínimo, mediana ou produto) (KITTLER et al., 1998; KUNCHEVA, 2004);
- *Threshold Selector*: é um algoritmo que escolhe um ponto como limiar de probabilidade da saída de um classificador selecionado. O limiar de ponto é estabelecido de modo que uma determinada medida de desempenho (por exemplo, medida-F, acurácia, precisão, revocação) seja otimizada. A partir do conjunto de treinamento, o desempenho é medido usando *hold-out* ou validação cruzada;
- *Stacking*: algoritmo proposto por Wolpert (1992), combina diversos classificadores utilizando um método de *Stacking*;
- *StackingC*: implementação mais eficiente do algoritmo *Stacking*, proposta por Seewald (2002);
- *Random SubSpace*: o classificador constrói várias árvores sistematicamente, com base em diferentes subconjuntos do espaço de características (diferentes subconjuntos de atributos originais) escolhidos aleatoriamente;
- *Random Committee*: o algoritmo utiliza classificadores que tem funcionamento aleatório como base. Cada modelo de classificação gerado é construído usando uma semente de número aleatório diferente (mas baseada nos mesmos dados). A previsão final é uma média das previsões geradas pelos modelos base individuais;
- *MultiScheme*: seleciona um classificador usando validação cruzada ou o desempenho no conjunto de treinamento. O desempenho para classificação é medido baseado na porcentagem correta, enquanto para regressão é o erro quadrático médio;

- *MultiBoostAB*: proposto por Webb (2000), é uma extensão para a técnica *AdaBoost*. O *MultiBoosting* pode ser visto como uma combinação do *AdaBoost* com "*wagging*";
- *LogitBoost*: algoritmo proposto por Friedman, Hastie e Tibshirani (1998), é um método estatístico baseado no *Boosting*, o qual cria modelos aditivos de regressão logística. Realiza classificação utilizando um regime de regressão como a base de aprendizagem, sendo indicado na classificação de problemas multiclasse;
- *Filtered Classifier*: executa um classificador após passar os dados por um filtro;
- *Dagging*: o algoritmo cria vários modelos a partir de diferentes subconjuntos, dos dados de treinamento, submetidos a cópias do classificador base. As previsões são feitas utilizando a média (TING; WITTEN, 1997);
- *Classification Via Regression*: o algoritmo realiza a classificação utilizando métodos de regressão, ou seja, transformam um problema de classificação em um problema de aproximação de função (FRANK et al., 1998);
- *Classification Via Clustering*: o algoritmo utiliza, para classificação, métodos de agrupamento. Para os algoritmos que utilizam uma configuração fixa de agrupamento, é necessário ter certeza que o número de grupos é igual ao número de classes no conjunto de dados;
- *Bagging*: a origem do termo *bagging* vem da expressão "*bootstrap aggregating*" (BREIMAN, 1996). O *Bagging* combina o método de *bootstrap*, de rearranjo, com o conceito de agregação. A ideia é combinar as classificações de múltiplos modelos ou do mesmo tipo de modelo, mas com diferentes conjuntos de dados para aprendizagem. A técnica de *Bootstrap* gera aleatoriamente várias amostras a partir da amostra original. O *Bagging* treina vários modelos a partir dessas amostras e realiza a combinação desses modelos;
- *Attribute Selected Classifier*: antes da classificação, é realizada a redução de dimensionalidade, por seleção de atributos, nos conjuntos de dados de treinamento e de teste;
- *AdaBoostM1*: proposto por Freund e Schapire (1996), é conhecido por *Boosting* Adaptativo, que, como o nome indica, é baseado no método de *Boosting* e pode ser utilizado em conjunto com um algoritmo de aprendizado de máquina. A cada iteração um algoritmo base é treinado utilizando uma versão do conjunto de dados. O algoritmo gera em cada passo uma distribuição sobre o conjunto de treinamento, dando um peso maior aos exemplos classificados incorretamente no passo anterior.

Logo, há diferentes versões ponderadas do conjunto de dados. Após um número determinado de iterações, o *Boosting* combina os diversos classificadores parciais, gerando um classificador único. Portanto, cada novo modelo é influenciado pelo desempenho do modelo anterior. O algoritmo *AdaBoost* lida com problemas binários, mas para classificação multiclasse tem-se a versão do algoritmo conhecida como *AdaBoostM1*;

- *Rotation Forest*: o algoritmo divide o conjunto de características em K subconjuntos, nos quais é aplicada a PCA separadamente. Os dados são transformados linearmente e utilizados para treinar o classificador de AD. Diferentes divisões do conjunto de características realizam diferentes rotações, assim, são obtidos diversos classificadores (RODRIGUEZ; KUNCHEVA; ALONSO, 2006).

2.4.5 Método de avaliação

Um método de avaliar os modelos de classificação é utilizar a análise ROC (do inglês, *Receiver Operating Characteristic*). Esse método é útil em problemas que possuem uma grande desproporção das classes ou em casos em que se deseja analisar o custo e o benefício entre diferentes resultados da classificação (BRADLEY, 1997; FAWCETT, 2006).

Para realizar a indução de um classificador, cujo aprendizado é supervisionado, é realizada a indução do algoritmo com instâncias rotuladas chamado de conjunto de treinamento. Para avaliar o desempenho do modelo treinado é utilizado um conjunto rotulado diferente, chamado de conjunto de teste.

Para um problema de classificação binária (isto é, as instâncias podem ser positivas ou negativas), uma maneira comum de apresentação e organização dos resultados é utilizar uma tabulação cruzada dos resultados da contagem entre a classe predita e a classe real. Essa tabulação é conhecida como matriz de confusão e pode ser observada na Tabela 2.

Tabela 2 – Matriz de confusão.

		Predito	
		V	F
Real	V	VP	FP
	F	FN	VN

Fonte: William T. Maruyama, 2015.

Quando um exemplo positivo é classificado como positivo é chamado de verdadeiro positivo (VP). Quando um exemplo negativo é classificado como positivo é denominado

de falso positivo (FP). Para a classe negativa segue-se a mesma ideia, ou seja, tem-se verdadeiro negativo (VN) e falso negativo (FN).

Além disso, algumas métricas podem ser calculadas utilizando os dados apresentados na Tabela 2. A proporção/taxa de acertos total da classificação é dada pela acurácia.

$$\text{Acurácia(ACC)} = \frac{\text{VP} + \text{VN}}{\text{Total da Amostra}} \tag{23}$$

O valor preditivo positivo (ou precisão): é a proporção de instâncias positivas corretamente classificadas como positivas (VP) em relação ao total de instâncias classificadas como positivas (VP+FP).

$$\text{Precisão} = \frac{\text{VP}}{\text{VP} + \text{FP}} \tag{24}$$

Sensibilidade (ou taxa de verdadeiro positivo, ou revocação): é a proporção de instâncias positivas classificadas como positivas (VP) em relação a todas as instâncias realmente positivas (VP+FN). Portanto, representa a capacidade do classificador detectar positivos.

$$\text{Revocação} = \frac{\text{VP}}{\text{VP} + \text{FN}} \tag{25}$$

Especificidade: é a proporção de instâncias negativas classificadas como negativas (VN) em relação ao total de instâncias negativas (VN+FP). Logo, indica a capacidade do classificador de detectar negativos.

$$\text{Especificidade} = \frac{\text{VN}}{\text{VN} + \text{FP}} \tag{26}$$

AUC (do inglês, *Area Under the Curve*): é a área sob a curva ROC. A curva ROC é uma representação gráfica para os valores de sensibilidade (eixo Y) e especificidade (eixo X). Huang e Ling (2005) argumentam que a AUC é uma medida melhor do que a acurácia na avaliação de diversos tipos de algoritmos.

Medida-F: é a média harmônica entre a precisão e a revocação. Representa o equilíbrio das duas métricas (BUCKLAND; GEY, 1994).

$$\text{Medida-F} = \frac{2}{\frac{1}{\text{precisão}} + \frac{1}{\text{revocação}}} \tag{27}$$

3 Revisão Sistemática

Para o entendimento do estado da arte na área de predição de relacionamentos optou-se pela realização de uma revisão sistemática.

3.1 Metodologia

Primeiramente foi realizada uma pesquisa exploratória sobre o tema *link prediction*, com o intuito de compreender os principais conceitos da área. Com base na pesquisa exploratória foi identificado que *link* ou *co-autorship prediction*, e *social network* ou *scientific collaboration network* como principais palavras-chaves relacionadas ao assunto. Posteriormente, por meio da metodologia de revisão sistemática foi criado o protocolo que define e formaliza os procedimentos seguidos nesta revisão (BIOLCHINI et al., 2005). A descrição do protocolo utilizado é apresentada nas próximas subseções.

A presente revisão sistemática tem como objetivo responder a seguinte pergunta: quais atributos (ou características) estão sendo utilizados na predição de coautorias em Redes Sociais Acadêmicas? Para responder a esta questão foram feitas pesquisas nas principais bibliotecas digitais científicas da área, as quais disponibilizam os trabalhos via Web. As bibliotecas digitais utilizadas neste trabalho são: IEEExplore Digital Library[1] e ACM Digital Library[2].

Com as bases e as palavras-chaves selecionadas foram criadas e submetidas as expressões e opções de busca em cada uma das fontes (Tabela 3). Para não restringir muito o resultado da busca, não foi considerado um período de publicação.

Tabela 3 – Chaves de busca utilizadas e condições utilizadas.

Fonte	**Expressão**	**Condições de filtragem**
ACM Digital Library	*((Abstract:"Link" OR Abstract:"co-authorship") AND (Abstract:"Prediction") AND (Abstract:"social network" OR Abstract:"scientific collaboration network"))*	Busca avançada, com utilização apenas do campo *abstract*
IEEExplore Digital Library	*((Abstract:"Link"OR Abstract:"co-authorship") AND (Abstract:"Prediction") AND (Abstract:"social network"OR Abstract:"scientific collaboration network"))*	Busca avançada, com filtro "*Metadata only*" ativo

Fonte: William T. Maruyama, 2015.

[1] http://ieeexplore.ieee.org/
[2] http://dl.acm.org/

Todos os artigos encontrados na busca foram avaliados e selecionados segundo os critérios de inclusão e de exclusão, que se seguem. Para aceitação do artigo, ele deve se enquadrar em todos os critérios de inclusão e nenhum de exclusão.

- Critérios de inclusão:
  1. Serão incluídos trabalhos completos publicados e disponíveis integralmente nas bases de dados científicas especificadas.
  2. Serão incluídos trabalhos que analisem Redes Sociais (não apenas acadêmicas).
- Critérios de exclusão:
  1. Serão excluídos trabalhos de estudos secundários.
  2. Serão excluídos trabalhos que não discutam os atributos que foram usados ou como foram usados para a predição de *links*.
  3. Serão excluídos trabalhos publicados que não estejam disponíveis integralmente nas bases de dados científicas especificadas.

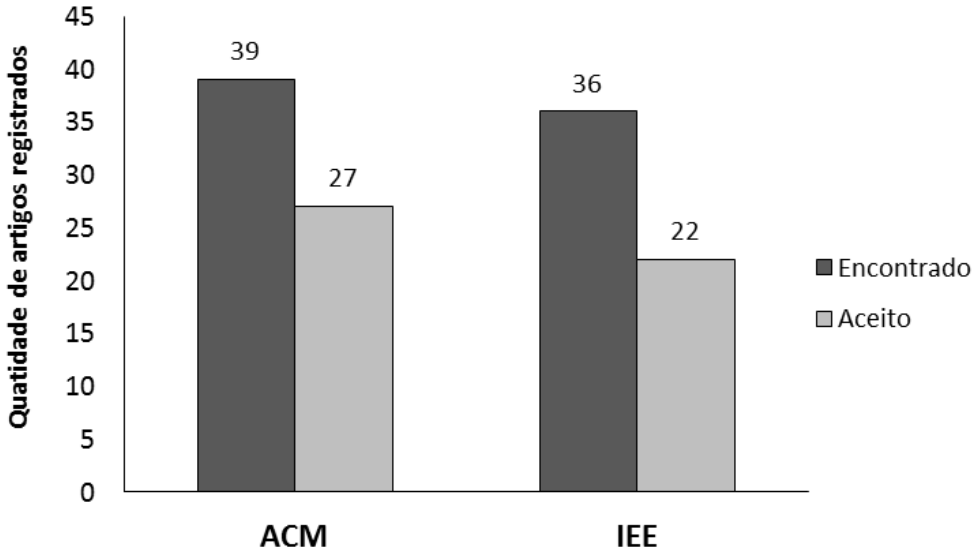
3.1.1 Condução

A submissão das expressões em cada um dos motores de busca das bibliotecas digitais foi realizada em Outubro de 2013. Foram encontrados: 39 artigos na ACM e 37 artigos na IEEE. Desses artigos, ocorreu apenas um caso de repetição. Portanto, 75 artigos foram analisados. Uma seleção inicial foi realizada a partir dos critérios de inclusão e exclusão, aplicados sobre a leitura dos resumos de cada artigo. Em caso de dúvida, o artigo foi lido na íntegra. Nesta etapa, 12 artigos encontrados na ACM e 14 na IEEE foram retirados pelos critérios de exclusão, portanto foi obtido 49 artigos que foram utilizados na etapa de extração. A Figura 4 apresenta a distribuição dos artigos nas respectivas bibliotecas digitais utilizadas.

3.1.2 Extração

Os 49 artigos incluídos nesta revisão foram lidos na íntegra e suas principais informações foram extraídas. Além dos dados bibliográficos, a Tabela 4 sumariza as informações extraídas de cada artigo, levando-se em consideração o foco do presente trabalho.

Figura 4 – Quantidade de artigos aceitos e rejeitados.



Fonte: William T. Maruyama, 2015.

Tabela 4 – Tabela de extração dos dados.

Referência	**Base de dados**	**Atributos utilizados**	**Domínio de aplicação**
Aiello et al. (2012)	Last.fm, aNobii	Informações do perfil do usuário	Predição de *links* de amizade
Almansoori et al. (2011)	Matriz com 24 encaminhamentos médicos	*Ethnicity* (E), *Professional Activity Match* (PAM), *Sum of Patients* (SofP), *Sum of Neighbors* (SofN), *Jaccard Similarity* (JS ou JC de *Jaccard Coefficient*)	Predição de *links* positivos entre médicos
Chang e Yao (2011)	Enron Email	*Singular value decomposition* (SVD), *Affinity measure* (AF)	Predição de *links* de trocas de e-mail
Chelmis e Prasanna (2012)	Serviço de microblog corporativo (semelhante ao Twitter)	*Shortest Distance*, CN, *Shared Vocabulary*, *SS_Uniform*	Predição *links* de intenção de comunicação
Corlette e Shipman III (2010)	Live Journal	AA, Coeficiente de Agrupamento Local	Predição de *links* de amizade (com efeito da abertura da rede)
Costa e Ortale (2012)	Small World network, Enron Email	Bayesian Hierarchical *Community-and-Role Model* (BH-CRM), *Latent Dirichlet Allocation for Graphs* (LDA-G)	Predição de *links* de interações de e-mails e citações
Cukierski, Hamner e Yang (2011)	Flickr	Katz, CN, AA, Cosseno, PA, *Bayesian Sets*, SVD *Features*, SimRank, EdgeRank, *Commute Time*, *Bounded Walk-PageRank*, *Maximum Flow*, *Betweenness Centrality*, *Core Number*, *Shortest Paths Histogram*, *Power Law Exponent*	Predição de *links* para separar relacionamentos reais de falsos

Continua na próxima página.

Referência	Base de dados	Atributos utilizados	Domínio de aplicação
Dong et al. (2012)	Epinions, Slashdot, Wikivote, Twitter	CN, AA, JC, PA, *ranking factor graph*, *out-degree*, *in-degree* e *all-degree*	Predição de *links* de interações em rede homogênea e heterogênea
Dong et al. (2011a)	PG, PB, Hep-th, Alex Arenas's Jazz, Alex Arenas's Email Network, Neural network of Elegans Network, US Air Network	CN, SA, LHN, SO, JC, HPI, HDI, PA, AA	Predição de *links* de interação em diversos tipos de redes
Dong et al. (2011b)	*Call Detail Records* (CDRs) de duas operadoras anônimas em uma cidade	CN, AA, JC, PA, HPI, HDI, SA, *Unweighted Random Walk* (URW), *Weighted call times random walk* (TRW), *Weighted call duration random walk* (DRW), *High-Performance Link Prediction* (HPLP), *Resource allocation based on weighted random walk* (RAURW), *Resource allocation based on weighted call times random walk* (RATRW), *Resource allocation based on weighted call duration random walk* (RADRW)	Predição de *links* de chamadas
Fire et al. (2011)	Academia, Facebook, Flickr, TheMarker, YouTube	*Vertex degree features*, *Vertex subgraphs features*, CN, *Total-Friends*, JC, *Transitive Friends*, PA, Katz, *Friends-measure*, *Opposite direction friends*, *Edge subgraphs edges number*, *Edge subgraphs components number*, SP	Predição de *links* faltantes de relacionamento em redes direcionadas e não direcionadas
Gao, Denoyer e Gallinari (2011)	Cond-mat, Gr-qc, Hep-ph, Hep-th	*Dependent Prediction method*, *Weighted Dependent Predition method*, CN, PA, Katz, *Nonnegative Matrix Factorization* (NMF), *Graph Nonnegative Matrix Factorization* (GNMF) e *Graph Regularized Joint Matrix Factorization* (GRJMF)	Predição de *links* temporal de coautoria
Gao, Denoyer e Gallinari (2012)	Live Journal, arXiv	NMF, *Mixed Membership Stochastic Blockmodels* (MMSB), *Multiplicative Latent Factor Model* (MLFM), *Generalized Latent Factor Model* (GLFM) e *Latent Factor BlockModel* (LFBM)	Predição de *links* de relacionamento social e de coautoria
Rodriguez e Rogati (2012)	LinkedIn	AA, CN normalizado, CA_e, AA_e	Predição *links* de conexão entre usuários após participarem do mesmo evento

Continua na próxima página.

Referência	Base de dados	Atributos utilizados	Domínio de aplicação
Guo e Guo (2010)	DBLP, TIG	*Merge Weighted Features Algorithm* (MWF)	Predição temporal de *links* de amizades e coautoria baseado em matriz para combinação de características
Hsieh et al. (2013)	LinkedIn, Enron Email, Wiki Talk	CN, AA, *Time overlap*, *Company size*, *Company average age*, *Company cluster coefficient*, *Node propensity*, *Join time difference*	Predição de *links* interação de usuário aderido há um tempo à rede (com *links*) e de usuário recém-aderido (sem *links*)
Huang et al. (2012)	Epinions	*Average Filling* (AF), JC, SimRank, SVD, *Matrix Completion* (MC), *Joint Manifold Factorization* (JMF)	Predição de *links* de confiança e desconfiança na rede social através da agregação de redes sociais heterogêneas
Jamali, Huang e Ester (2011)	Flixster, Epinions	*Generalized Stochastic Blockmodel* (GSBM) e *Mixed Membership Stochastic Blockmodel* (MMB)	Predição de *links* entre usuários em uma Rede Social de Avaliação
Kamei et al. (2012)	@cosme	JC, *Cosine Similarity* (CS ou SA de Salton), Modelo probabilístico porposto com características latentes	Predição de fan-*links* faltantes com base nos dados observados de atividades do usuário
Kunegis, Preusse e Schwagereit (2013)	Epinions, Slashdot	JC, AA, *Exponencial kernel*, PR *product*, CN, *Paths of lenght three*, similaridade por cosseno, PA e PR condicional	Predição de *links* negativos em rede sociais
Kuo et al. (2013)	Foursquare, Twitter, DBLP	*User friendship* (UF), *Item ownership* (IO), *Category popularity* (CP), BC, JC, PA, *Attractiveness* (AT), PageRank *with Priors* (PRP), AT-PRP, *Infer* e *Learn*	Predição de *links unseen-type* em uma rede heterogenea
Lerman et al. (2012)	Digg, Twitter	CN, JC, AA, CS, *Attention-limited Conservative Metric* (CS_AL), *Non-Conservative Proximity* (NC) e *Attention-Limited Non-Conservating Proximity* (NC_AL)	Predição de *links* de atividade
Leroy, Cambazoglu e Bonchi (2010)	Flickr	CN, Katz, *rooted* PR	Predição de *links* entre os usuários em *cold start*
Liben-Nowell e Kleinberg (2003)	astro-ph, cond-mat, gr-qc, hep-ph, hep-th	CN, JC, AA, PA, *rooted* PR, Katz, *Hitting time*, SimRank, Distância no grafo. Meta-abordagens: *Low-rank approximation*, *unseen bigrams* e *clustering*	Predição temporal de *links* de coautoria

Continua na próxima página.

Referência	Base de dados	Atributos utilizados	Domínio de aplicação
Lin, Yun e Zhu (2012)	Interactome, USAir, C. elegance, CGScience	CN, AA, RA, *Weighted* CN (WCN), *Weighted* AA (WAA), *Weighted Resource Allocation* (WRA), *BenefitRanked* CN (BrCN), *BenefitRanked* AA (BrAA) e *BenefitRanked* RA (BrRA)	Predição de diversos tipos de *links* faltantes em redes ponderadas
Lu et al. (2010)	Hep-th, CiteSeer, SIAM	Katz *single source* (Katz-S), Katz *all source* (Katz-C), *Truncated* Katz *single source* (tKatz-S), *Truncated* Katz *all source* (tKatz-C), *Supervised Learning single source* (SL-S), *Supervised Learning pure color path* (SL-P), SL-P com L1, *Supervised Learning hybrid color paths* (SL-H), SL-H com regularização L1 e SL-H com regularização hierárquica estruturada (HS)	Predição de *links* de coautoria
Makrehchi (2011)	Informação bibliográfica de publicações em 20 domínios científicos coletados da Web	*Latent* Dirichlet *Allocation* (LDA) com Katz, LDA com SP, *Bag-Of-Words* (BOW) e *Latent Semantic Indexing* (LSI)	Predição de *links* de coautoria, a partir da semelhança entre resumos em coautoria entre os autores
Nie et al. (2012)	Wikipedia, Slashdot	CN, SVD, *Fixed Point Continuation* (FPC), *Accelerated Proximal Gradient* (APG), Método proposto pelos autores	Predição de *links* faltantes de interação entre usuários
Perez, Biregah e Lemercier (2012)	Um conjunto de redes sociais (Address Book, Twitter, Google+ e Facebook) extraído de iPhones e um conjunto de contatos (amigos e não amigos) extraídos do Facebook	CN, SA, JC, HPI, HDI, LHN, PA, AA, RA, WRA	Predição de *links* para detecção de contatos ilegítimos
Quercia e Capra (2009)	Parte dos dados do projeto *Reality Mining* do MIT	SP, PR, HITS, KmarkovChain	Predição de *links* para recomendar amigos com base na proximidade do celular
Sa e Prudencio (2011)	DBLP	CN, JC, PA, PD, RA, LP, Coeficiente de agrupamento local	Predição de *links* temporal de coautoria em uma rede ponderada

Continua na próxima página.

Referência	Base de dados	Atributos utilizados	Domínio de aplicação
Shin, Si e Dhillon (2012)	Flickr, LiveJournal, MySpace, Epinions	PA, AA, *Random Walk with Restart* (RWR), CN, Katz. Outros métodos de aproximação: *Eigen-decomposition*-CN (EIG-CN), *Clustered Low Rank Approximation*-CN (CLRA-CN), *Multi-Scale Link Prediction*-CN (MSLP-CN), EIG-Katz, CLRA-Katz e MSLP-Katz	Predição de *links* explorando diferentes escalas de aproximação para redes sociais de grade escala
Soares e Prudencio (2012)	Hep-th, Hep-lat	Métricas de similaridade: AA, PA, CN, JC. Combinados com os métodos: *Moving Average* (MA), *Average* (Av), *Random Walk* (RW), *Linear Regression* (LR), *Simple Exponential Smoothing* (SES), *Linear Exponential Smoothing* (LES)	Predição de *links* de coautoria considerando séries temporais
Song et al. (2009)	Digg, Flickr, LiveJournal, MySpace, YouTube, Wikipedia	PA, PRP, CN, AA, Katz, Distância no grafo (*Graph distance*, GD)	Predição de *links* de relacionamentos em redes sociais de alta escala
Song et al. (2012)	Flickr, LiveJournal, MySpace	Aprendizagem espectral com *Clustered Spectral Graph Embedding* (CSGE), Katz com *Spectral Graph Embedding* (SGE), CN	Predição de *links* e *links* faltantes de relacionamento
Steurer e Trattner (2013)	Second Life (posição dos usuários), My Second Life	CN, JC, AA, PA, *Common Groups* (GC), *JC for Groups* (GJC), *Common Interests* (IC), *JC for Interests* (IJC), *Common Regions* (RC), *Regions Seen Concurrently* (RS), *Observations Together* (RO), *Physical Distance*, *Days Seen*	Predição de *links* de interação entre usuários, através de análise dos dados de posição e da rede social
Tian et al. (2010)	Facebook, Chamadas de celulares (CALL)	Link *trend*, Número de interações totais, Número de recentes interações, Tempo da última interação, número de intervalos de tempo ativo, CN, JC, CN ativos, Número total de amigos, Número total de interações	Predição de *links* para reconexão de *links* em redes de interação social

Continua na próxima página.

Referência	Base de dados	Atributos utilizados	Domínio de aplicação
Tylenda, Angelova e Bedathur (2009)	DBLP, astro-ph	Versões de PR e AA padrões e ponderadas por Ano da mais recente colaboração (last), Número de colaborações (count), Número mínimo de coautores (min. coauth). ME, TME *avg.*, *exp.*, TME *avg. lin.*, TME *avg. sqrt.*, TME *sum lin.*, Distância (dist), JC, CN, *last count*, *count last*, *min. coauthors*, *dist. last count*, *dist. count last*, *dist. min. coauth*, ordenação por *count last*, ordenação por *last count*	Predição de *links* de coautoria, novos e repetidos
Valverde-Rebaza e Lopes (2012)	Twitter	*Within And Inter Cluster* (WIC), CN, AA, JC, RA, PA	Predição de *links* de seguidores no Twitter
Vasuki et al. (2010)	Orkut, Youtube	tkatz, SVD	Predição de *links* para recomendação de comunidades
Vasuki et al. (2011)	Orkut, Youtube	tKatz (aplicado em dois grafos), tKatz com *latent factor model* (tKatz-LFM), tKatz com *common subspace model* (tKatz-CS), tKatz com *clustered latent factor model Equation* (tKatz-LFM-c)	Predição de *links* para recomendação de grupos ou comunidades em redes de grande escala
Wang et al. (2011a)	CDRs	Katz, AA, CN, JC, *Spatial Cosine Similarity*, *Weighted Spatial Cosine Similarity*, *Extra-role Co-Location Rate Weighted*, *Weighted Co-Location Rate Common* e *Co-Location Rate*	Predição de *links* de chamadas com medidas de mobilidade
Wang et al. (2011b)	CORA	*Dynamic Relational Topic Model* (dRTM), RTM	Predição de *links* de citação com um modelo capaz de lidar com *links* ruidosos
Wang, Satuluri e Parthasarathy (2007)	DBLP, Genetics, Biochemistry	*Approximate Katz measure* (aKatz), *Co-occurrence probability*, AA, PA	Predição de *links* de coautoria utilizando um novo modelo probabilístico em rede de coautoria
Xia et al. (2012)	Internet Movie Database	CN, JC, AA, *User-based collaborative filtering* CF_u, *Item-based collaborative filtering* CF_i, PA, Katz, *Minimum Description Length* (MDL), *Absent Links* (AL), RWR	Predição de *links* entre diretor e ator de filmes com métodos adaptados de métodos tradicionais baseados em vizinhança para redes sociais bipartidas

Continua na próxima página.

Referência	Base de dados	Atributos utilizados	Domínio de aplicação
Yin, Hong e Davison (2011)	Twitter	PropFlow, Katz, JC, AA, CN, PA, Matriz de fatoração	Predição de *links* de seguidores em uma rede híbrida
Yu et al. (2011)	Geraram quatro conjuntos de dados sintéticos e o conjunto de dados MIT Reality Mining Project	Same Edge, Global Positioning System Similarity (GPSSim), RWR, *GeoFriends Recommendation Framework* (GEFR)	Predição de *links* para recomendação de amigos em uma rede social *cyber-physical*
Zhang, Zhai e Wu (2013)	Sina Microblog	*Exponential random graph model* (ERGM), JC, Katz	Predição de *links* de relacionamento nas comunidades de um microblog
Zhong et al. (2013)	Tencent, SinaWeibo, Xiaone, Facebook, Twitter, Github, Stackoverflow, Epinions	*Time-evolving Composite Network Models* (ITCom), *Mixed Membership Stochastic Blockmodels* (MMSB), *dynamic Mixed Membership Stochastic Blockmodels* (dMMSB), *Nonparametric Metadata Dependent Relational Model* (NMDR), *dynamic Infinite Relational Model* (dIRM), *Tensor Factorization* (TF)	Predição temporal de *links* de interação e amizade entre usuários

Fonte: William T. Maruyama, 2015.

3.2 Resultados

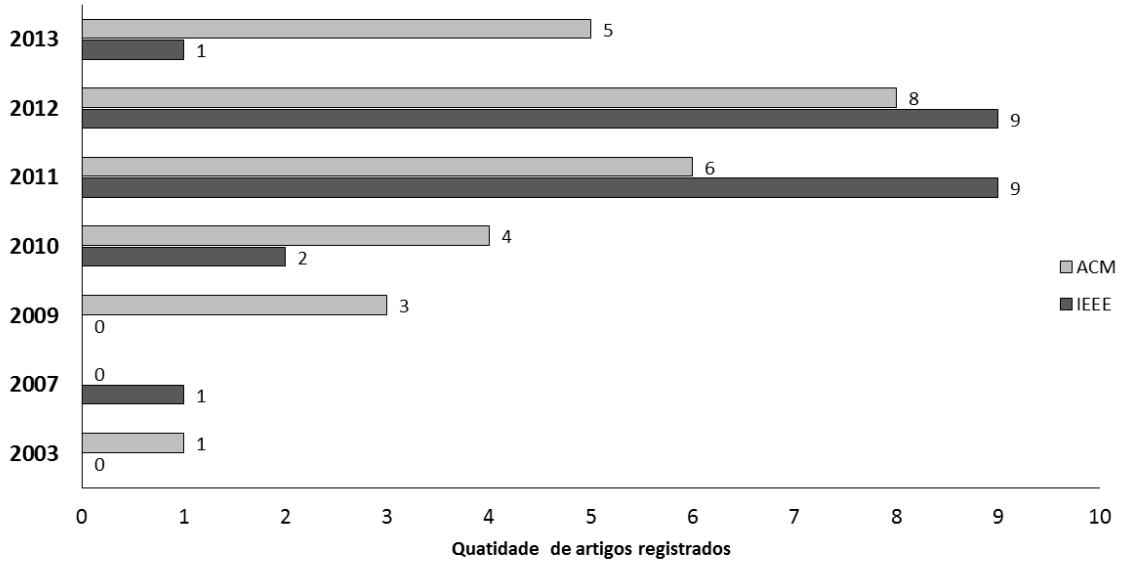
A partir dos artigos incluídos, foi observado que as publicações sobre este assunto são recentes. Esta informação pode ser verificada na Figura 5.

De 2010 a 2011 ocorreu um aumento de mais de 70% nas publicações sobre assunto. Além disso, é nos últimos três anos que se concentra cerca de 77% das publicações. Portanto, é possível observar que este assunto está em alta na comunidade científica.

Com os mesmos artigos, foi possível realizar uma análise quanto à distribuição geográfica dessas publicações (Figura 6) tomando como base os dados de localização do primeiro autor. Os resultados mostraram que as pesquisas nessa área se concentram nos EUA, com 21 publicações, seguido pela China com 10. O Brasil localiza-se em quarto lugar com 3 publicações, na qual a Universidade Federal de Pernambuco e a Universidade de São Paulo, são as instituições onde foram encontradas as publicações sobre o assunto.

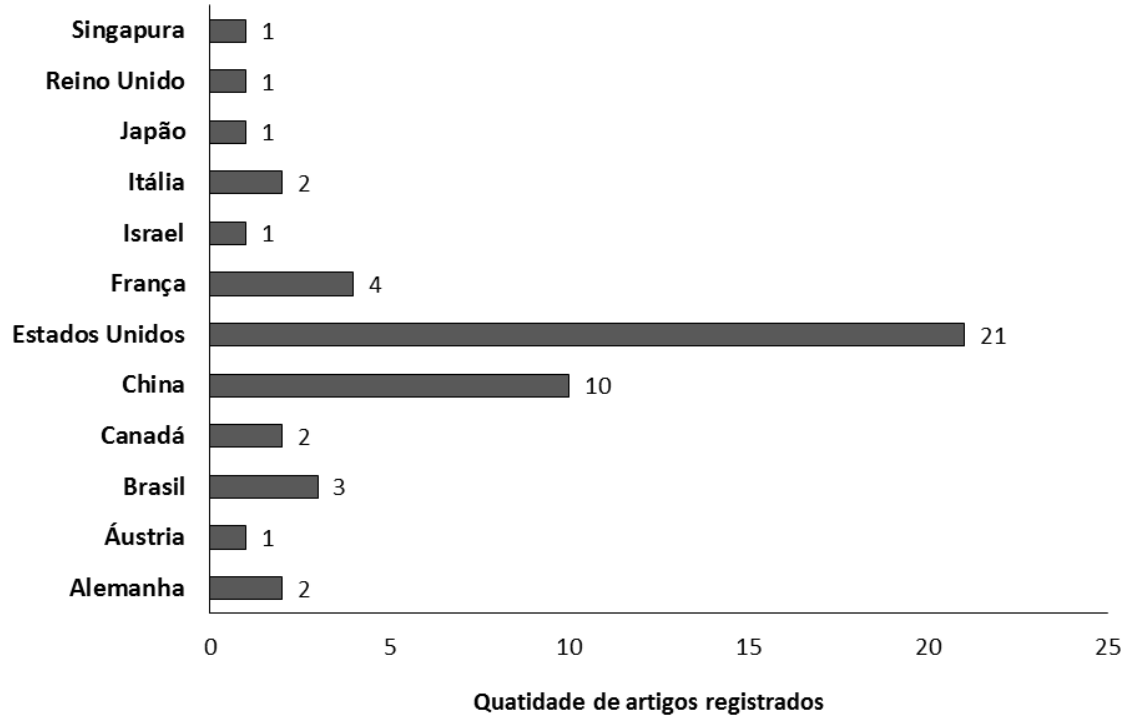
Sobre os conjuntos de dados utilizados nos 49 artigos incluídos nesta revisão, foram registradas 57 fontes de dados diferentes (Figura 7). Para tal análise, deve-se levar em

Figura 5 – Quantidade de publicação sobre o tema ao longo dos anos.



Fonte: William T. Maruyama, 2015.

Figura 6 – Distribuição geográfica das publicações sobre o assunto.



Fonte: William T. Maruyama, 2015.

consideração que um artigo pode ter utilizado mais de uma base de dados e cada repetição foi contabilizada. A quantidade de fontes e sua distribuição nos artigos, principalmente dos 38 conjuntos de dados utilizados uma única vez em todo levantamento, demonstram que há grande variedade de domínios nos quais a predição está sendo realizada. Isso mostra que esse

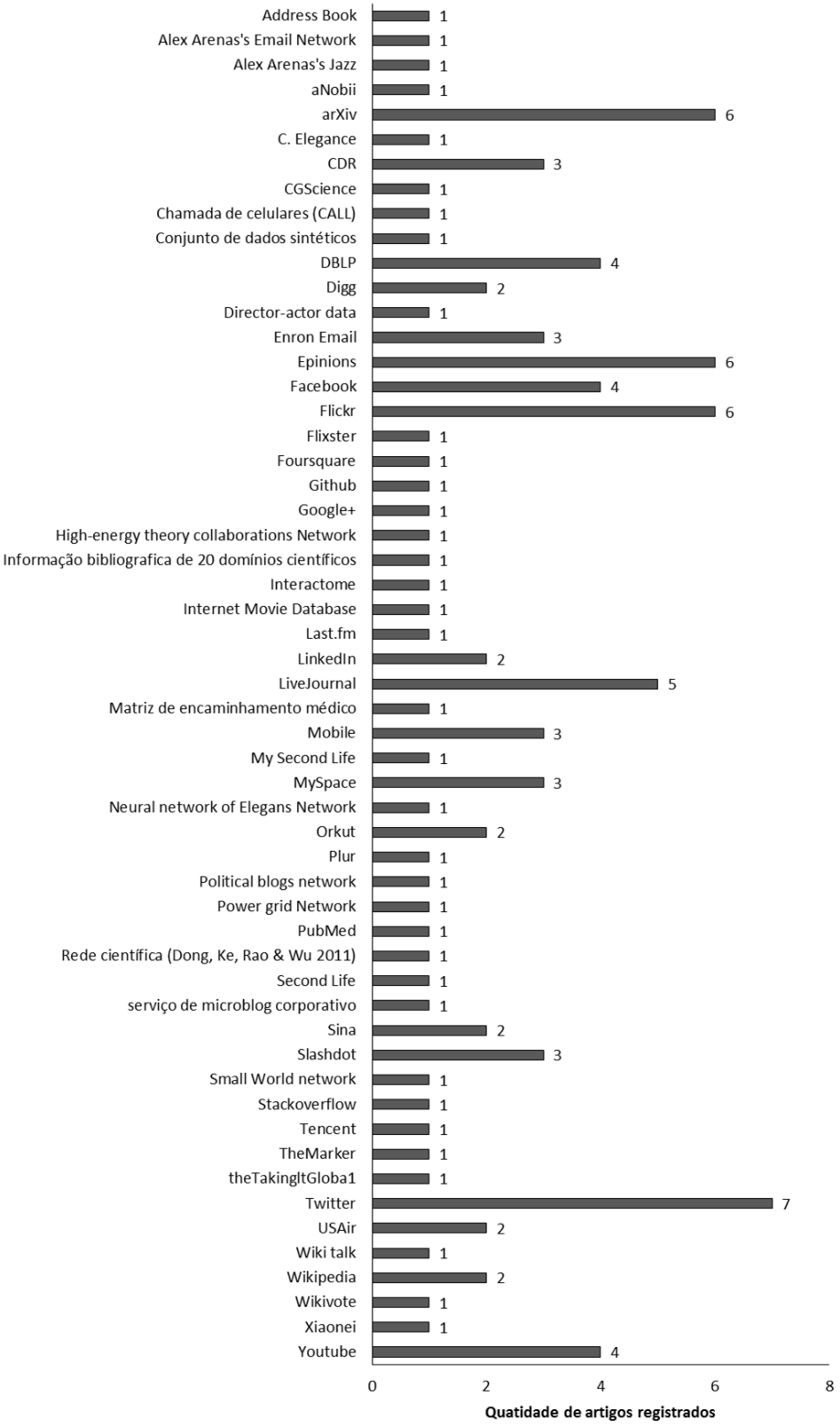
tema é amplo e pode ser aplicado aos mais diversos conjunto de dados, desde redes sociais *online* de amizade, passando por redes sociais acadêmicas e profissionais, abrangendo até a telefonia móvel. Os conjuntos de dados mais utilizados foram os provenientes do Twitter (7 registros), seguido de arXiv, Epinions e Flickr (6 registros).

Dentre todos os 49 artigos, o primeiro publicado (e mais citado) sobre a predição de *links* foi o de Liben-Nowell e Kleinberg (2003) no qual os autores propuseram a predição de arestas (*links*) futuras com base nas arestas atuais, utilizando diversas medidas de proximidade (atributos) de nós em uma rede para realizar predição dos *links* (CN, JC, AA, PA, Katz, Hitting time e SimRank). Para tal, os autores utilizaram conjuntos de dados do arXiv, para a realização de predição de *links* de coautoria. Como resultado, eles concluíram que o atributo Katz e as variantes utilizadas apresentaram bom desempenho na maioria dos subconjuntos de dados, sendo os melhores resultados obtidos em três dos cinco subconjuntos. Além disso, segundo os autores, os atributos simples como CN e AA apresentaram resultados satisfatórios. A partir desta publicação, a maioria dos artigos incluídos na presente revisão propõe novos atributos ou técnicas para predizer a formação de uma nova aresta entre os inúmeros nós de uma rede.

Com a mesma fonte de dados (arXiv) para construção da rede, mas não necessariamente com os mesmos dados, há outros 5 artigos que foram incluídos na presente revisão. Dentre eles, incluem-se a publicação de Gao, Denoyer e Gallinari (2011), na qual os autores propuseram um modelo unificado de múltiplas informações da rede para predizer *links* de coautoria em função do tempo. Essas informações são de três tipos: da estrutura global da rede, o conteúdo dos nós e as informações de proximidade nos grafos para capturar os padrões de evolução ao longo do tempo das ligações nas redes. Utilizando quatro conjuntos de dados do arXiv entre 1992 e 2002, os resultados apresentados demonstram, segundo os autores, que o método proposto é eficiente em vários conjuntos de dados, podendo, de acordo com os valores da área sob a curva (AUC), superar os métodos tradicionais para predição de *links* temporais. Os autores comentam também a possibilidade do uso da solução proposta em redes de larga escala.

Tylenda, Angelova e Bedathur (2009) utilizaram o conjunto de dados de astrofísica do arXiv para desenvolver um método de predição de *links* em grafos, com a incorporação da informação temporal contida na evolução das redes sociais, ampliando um modelo probabilístico tradicional. Esta proposta incorpora pesos nas arestas derivadas de características temporais nos métodos de predição de *links*. Os resultados mostram, segundo os autores,

Figura 7 – Uso dos diferentes conjuntos de dados registrado nos 49 artigos incluídos.



Fonte: William T. Maruyama, 2015.

que o conhecimento das interações temporais entre usuários melhora significativamente a acurácia da predição de novos *links*.

Com o intuito de predizer novos *links* considerando o comportamento dos relacionamentos conforme a série temporal, Soares e Prudencio (2012) utilizaram duas seções do arXiv. As duas seções consistem no Hep-th (1991 a 2010) e Hep-lat (1993 a 2010). A ideia básica é a construção de séries temporais para cada par de nós não conectados, usando um *score* de similaridade calculado por uma métrica topológica. Um modelo de previsão é então utilizado, a fim de prever o valor seguinte da série. Esse valor é o *score* final do par de nós a ser usado pelos métodos de predição de *links*, testado conforme uma abordagem supervisionada e não supervisionada. Segundo os autores, a abordagem supervisionada foi melhor em todos os modelos de previsão em relação à abordagem não supervisionada, mas este trabalho ainda apresenta limitações quanto ao número de redes utilizadas nos experimentos e seus domínios.

Com outro enfoque, Lu et al. (2010) também utilizaram conjuntos de dados do arXiv (Hep-th de 1992 a 2003), juntamente com os dados do CiteSeer (1993 a 2003) e do SIAM (1999 a 2004) em sua pesquisa, que tinha como objetivo propor um novo método de predição de relacionamentos de coautoria, citação e referência, respectivamente. Utilizaram uma abordagem supervisionada, múltiplas fontes e observações do histórico da rede. Segundo os autores, os resultados experimentais confirmam que a precisão da predição de *links* na dinâmica da rede de uma série temporal é melhorada utilizando a técnica supervisionada e de múltiplas fontes auxiliares de informação, provenientes de outras redes.

Uma abordagem alternativa para a predição de coautorias foi encontrada em Makrehchi (2011), na qual o autor propõe uma abordagem que faz predição baseada na extração de temas ocultos a partir de dados de texto. A ideia é construir uma rede social a partir de documentos de texto, extraindo semelhanças semânticas entre os nós que estão associados aos documentos. O objetivo foi prever as coautorias a partir das semelhanças entre os resumos dos autores. Para tal, o autor utilizou a informação bibliográfica de artigos e resumos de 20 domínios científicos (Acústica, Dermatologia, Microbiologia, Estatística, Zoologia, entre outros) coletados da Web. Conforme o autor, os resultados obtidos no teste dessa abordagem de predição de *links* em redes sociais possui alta precisão.

Nessa linha de pensamento de utilização de mais de uma rede como complemento para a predição de *links*, os trabalhos de Huang et al. (2012), Kuo et al. (2013), Dong

et al. (2012) utilizaram combinações de informações, que podem ser entendidas como heterogêneas. No primeiro trabalho, os autores propõem o uso do método JMF (*Joint Manifold Factorization*) para prever *links* de confiança e desconfiança na rede social por meio da agregação de redes (de domínio alvo e de informação auxiliar). De acordo com os autores, os experimentos realizados comparando o JMF com outros métodos demonstraram a eficácia do método proposto. Essa mesma proposta de agregação foi realizada por Kuo et al. (2013) que, utilizando estatística agregativa e modelo probabilístico não supervisionado, desenvolveram um modelo de fator de grafo (o *Factor Graph Model with Aggregative Statistics*) com três camadas de variáveis aleatórias (candidato, atributo e contagem), que combinadas formam uma rede heterogênea, para inferir a existência de *unseen-type link*. Este tipo de *link* faz parte de um conjunto invisível numa rede heterogênea e influencia na dinâmica da mesma. Para tal, os autores testaram seu método com cada um dos conjuntos de dados do Foursquare, Twitter, Plur e DBLP. De acordo com os autores, os resultados comparativos demonstram que o método proposto, com a variação chamada de LEARN, obteve melhor desempenho em todos os conjuntos de dados. Portanto, eles concluíram que as informações heterogêneas podem ser combinadas em um *factor graph*.

Em Dong et al. (2012) é proposto um método chamado *Ranking Factor Graph* (RFG) que, em redes obtidas de diferentes fontes, é adaptado e denominado de *transfer-based RFG* (TRFG). O TRFG combina as informações de diferentes redes e sua ideia está baseada nos princípios de homofilia, que sugere que usuários com características semelhantes tendem a se associar e a compartilhar um vínculo positivo em uma rede. Com base nisso, é possível que em diferentes redes, os mesmos usuários tendem a se associar também. Conforme os autores, os métodos propostos tiveram bons resultados, ao serem comparados com outros métodos considerados básicos. Eles comentam também que a escala das redes sociais *online* está crescendo a uma taxa exponencial e o mecanismo de formação de vínculo, isto é, as interações humanas, é ainda pouco explorado.

Baseado no mesmo princípio de homofilia, Lerman et al. (2012) e Aiello et al. (2012) desenvolveram suas pesquisas. No primeiro trabalho, os autores comentam que as pessoas que estão mais próximas em uma rede são mais propensas a realizar ações semelhantes, do que pessoas que estão mais distantes. Os autores utilizaram Twitter e o Digg como base de dados para suas pesquisas por possuírem duas características de natureza diferente nas interações em mídias sociais: em uma é possível estabelecer uma relação de interação a partir de uma pessoa para muitas e, na outra, há a limitação da

capacidade do usuário de responder aos estímulos recebidos, chamada Atenção Limitada (AL). Os autores calcularam a proximidade para obter o grau do quanto as pessoas estão próximas e comparam diferentes métricas de proximidade tradicionais (CN, JC, AA e CS) e introduzidas (CS_AL e NC_AL) na tarefa de predição de *links*. Os autores mostram que as métricas de proximidade estrutural levando em conta a atenção limitada, obtém melhores resultados em precisão e revocação, pois argumentam que com a abordagem proposta representam a natureza da comunicação da rede.

Aiello et al. (2012), por sua vez, têm sua publicação dividida em diferentes análises, dentre as quais a homofilia e a predição de *links*. Para testar a hipótese de que a presença de laço social pode ser prevista com base apenas na similaridade tópica entre os nós foram utilizadas duas bases de dados. Os nós foram comparados par a par quanto à similaridade tópica (isto é, os atributos conforme os dados do perfil dos usuários), para ter seus resultados ranqueados e, depois, comparados com outras métricas (atributos baseados nas características da rede). De acordo com os autores, houve uma forte correlação entre vinculação social e atividade do usuário, mostrando um fenômeno de homofilia entre eles, que pode ser utilizado para fortalecer os resultados na predição de *links*. Além disso, CN mostrou-se como um bom preditor de *links* sociais, tendo desempenho melhor que todos os outros métodos testados.

Utilizando oito bases de dados, parte delas de redes sociais acadêmicas, Dong et al. (2011a) visam provar a eficiência da métrica proposta, baseada nos padrões regulares encontrados por eles em experimentos prévios. Esta métrica, chamada de *Degree Exponent Change*, leva em consideração o grau do expoente de vizinhos em comum nos cálculos de similaridade, para prever *links* entre dois nós. Segundo os autores, os resultados de AUC são satisfatórios na predição pela métrica proposta, entretanto são necessários mais estudos em diferentes redes com diferentes características topológicas e citam a necessidade da realização de um estudo aprofundado na estrutura interna da rede.

Sa e Prudencio (2011) também trabalharam com redes sociais acadêmicas. Com o intuito de investigar a relevância do uso de pesos nas ligações (arestas) para melhorar a predição de *links* na abordagem supervisionada e na não supervisionada, os autores adaptaram uma métrica para esse fim. Isso porque esses pesos expressam a intensidade das relações. Os autores utilizaram o conjunto de dados do DBLP, dividido em três subconjuntos: não ponderada caso os dois autores já foram coautores de um mesmo artigo, ponderada de acordo com o número total de trabalhos em que o par de autores foi coautores

e ponderada pela contribuição dos autores em seus trabalhos de coautoria. Conforme mostram os autores, em quase todas as comparações entre as redes, a rede não ponderada obteve um desempenho inferior em relação a, pelo menos, uma das redes ponderadas. Os autores concluem que, embora estes resultados não sejam conclusivos, é possível realizar melhorias no desempenho da predição de *links* ao se considerar os pesos de cada ligação. Outro trabalho que se utilizou de redes ponderadas foi o de Guo e Guo (2010), no qual os autores utilizam uma matriz para combinar a ponderação das características atribuídas, características topológicas e característica temporal de uma rede. Os autores utilizaram duas bases de dados (uma rede de coautoria e uma rede social de interação *online*) para testar o algoritmo que faz essa combinação, o *Merge Weighted Features* (MWF). Segundo os autores, o método proposto indicou uma melhora no desempenho na predição de *links*, isso porque houve a atribuição de pesos às características importantes. Eles também comentam que este método pode ser utilizado em outras áreas, como redes de amizades, de interação gênica, dentre outras.

Lin, Yun e Zhu (2012) também utilizaram uma rede ponderada para o delineamento de uma medida de similaridade para predição de *links*. Foi utilizada uma combinação do modelo de Cadeias de Markov com a teoria de laços fracos, para obter informações de nós vizinhos e em seguida calcular o valor *BenefitRank* de cada nó na rede ponderada. A ideia é que o *BenefitRank* de um nó represente implicitamente a quantidade de informação coletada de seus vizinhos próximos, no qual um valor alto significa muitos vizinhos. De acordo com os autores, combinar Cadeias de Markov com a teoria de laços fracos pode efetivamente identificar os papéis das diferentes ordens de vizinhos e alcançar maior precisão.

Algumas pesquisas como a de Rodriguez e Rogati (2012) são mais complexas ao considerar não só a interação *online*, mas também a interação *off-line*, através de encontros sociais ou profissionais, entre os usuários. Com o intuito de mostrar como os eventos profissionais e encontros sociais no mundo real se relacionam com a dinâmica temporal e evolução de uma rede profissional, os autores concluíram que novos *links* são realizados em curto período após a data do evento e que sua predição é mais eficiente nesse período. Além disso, a conexão entre nós distintos possui influência dos nós em comum que ambos compartilham.

Uma pesquisa considerando a tecnologia de telecomunicação foi realizada por Quercia e Capra (2009), a qual propõe o *FriendSensing*, que sugere amigos automati-

camente para os usuários de redes sociais móveis. Para isso os algoritmos são baseados em proximidade geográfica, conforme o alcance do Bluetooth do dispositivo móvel. O experimento foi realizado entre 2004 e 2005, utilizando celulares com Bluetooth habilitado de 96 funcionários e estudantes do campus do *Massachusetts Institute of Technology* (MIT), participantes do *Reality Mining Project*. A rede social foi criada a partir de informações contidas em registro das ligações realizadas e mensagens de textos enviadas. Segundo as autoras, as informações não geográficas devem ser consideradas, pois a amizade depende também de uma ocupação profissional semelhante, da preferência cultural ou do grupo social que o individuo frequenta.

Utilizando-se também do *Reality Mining Project*, Yu et al. (2011) propuseram um método chamado GEFR (*Geo-Friends Recommendation Framework*), que tem como objetivo recomendar amigos geograficamente relacionados em redes sociais. Esse método extrai informações de padrões interessantes e discriminativos a partir de uma grande quantidade de dados brutos de GPS (*Global Positioning System*) e combina com informações estruturais da rede social, construindo um padrão de acordo com uma rede de informações heterogêneas e definindo uma matriz de probabilidade de transição para descrever todas as probabilidades de transição de um conjunto de arestas. Aplicando o método de RW nessa rede de informações, *links* relevantes entre diferentes nós podem ser estimados e potenciais *geo-friends* podem ser recomendados para um usuário específico. Conforme os autores, os resultados de precisão e revocação foram melhores em GEFR do que nos outros métodos, mas esta melhora não é estatisticamente significativa.

Outras pesquisas realizadas com sistema móveis (*mobile*) são os trabalhos de Dong et al. (2011b), Wang et al. (2011a), Perez, Birregah e Lemercier (2012). No primeiro, os autores têm como intuito modelar as múltiplas facetas de uma vida digital a partir dos dados disponíveis nos *smartphones*. Eles utilizaram o conjunto de dados provenientes de diferentes aplicativos móveis e a rede de amigos do Facebook do usuário e propuseram uma função que formaliza as conexões intercamadas (entre as redes sociais) do modelo, chamando-o de *MultiLayer model*. De acordo com os autores, a precisão desse método apresentou-se eficiente, principalmente para a detecção de contatos ilícitos por predição de *links*, contribuindo para um quadro de apoio de prevenção de vazamento de dados.

No segundo trabalho, Dong et al. (2011b) utilizam o conjunto de dados de duas operadoras de telefonia celular de uma cidade. Os autores aplicaram para medir a similaridade na predição de *links* o *Resource Allocation*, que é um conceito proveniente da física

teórica. Foi definido o recurso de propriedade de cada nó com atributos das chamadas, que representam os níveis de atividade, tais como a frequência e duração das chamadas, para testar sua proposta de método híbrido para predição não supervisionada de *links*. Essa metodologia é baseada nas metodologias do *Random Walk*, que orienta o processo de *Resource Allocation*, combinando as informações topológicas das redes com os atributos de nós e arestas. Segundo os autores, a metodologia proposta obteve resultados de AUC que superaram outras abordagens não supervisionadas.

Em Wang et al. (2011a), o foco principal do trabalho é explorar o poder preditivo de mobilidade individual comparado e combinado com atributos topológicos. Para tal, utilizaram as trajetórias e os padrões de comunicação de uma base anônima de um país, cujos dados são obtidos de CDR (*Call Detail Record*). Segundo os autores, os resultados demonstram que a mobilidade tem forte influência na predição de *links*, conforme a correlação entre a semelhança nos movimentos dos indivíduos, suas conexões sociais e a força das interações entre eles. Combinando as medidas de mobilidade e de rede, os autores mostraram que a precisão na predição pode ser significativamente melhorada com aprendizado supervisionado.

Neste contexto de conjunto de dados massivos, as pesquisas utilizando redes de larga-escala também foram encontradas dentre os artigos incluídos. Em alguns destes trabalhos, os autores combinam algumas técnicas com atributos tradicionalmente conhecidos. Este é o caso de Song et al. (2009), no qual os autores desenvolveram duas novas técnicas, o *proximity sketch* e o *proximity embedding*, para estimar medidas (atributos) de proximidade em redes de larga escala. Os autores testaram essas medidas utilizando cinco redes de larga escala e obtiveram como resultado que essas medidas foram eficazes para predição de *links*, variando significativamente entre diferentes redes sociais. Além disso, a combinação das medidas de proximidade utilizadas com a árvore de decisão produziu uma melhor precisão na predição. Corlette e Shipman III (2010), por sua vez, estudaram a dinâmica dos *links* entre os usuários ao longo do tempo. Para isso, seguiram alguns usuários a partir do momento em que entram em uma rede de larga escala até 10 meses após a adesão e examinaram o efeito da aplicação de predição de ligação. A ideia foi analisar a dinâmica de vinculação ao longo do tempo entre os usuários e os efeitos que a entrada do usuário na rede tem sobre a predição de novos laços. Foi utilizada uma abordagem supervisionada, com o classificador Naive Bayes e com duas métricas como atributos. Conforme os autores, o experimento mostrou que os resultados da predição são melhores logo após a entrada

do usuário na rede e que a precisão e a revocação dos resultados diminuem quanto mais tempo os usuários estão na rede.

Diferentemente, Vasuki et al. (2011) abordaram a recomendação de grupos e comunidade para os usuários com informações da rede de amizades dos mesmos. Para tanto, os autores propuseram dois métodos que podem ser utilizados em redes de larga escala. Os métodos são o *commom subspace approximation* e *clustered low rank approximation*. Comparando o desempenho dos métodos propostos com outros métodos tradicionais em dois conjuntos de dados de redes sociais, os autores constataram que os métodos baseados em grafos de proximidade foram mais eficazes.

Outra pesquisa realizada com rede de larga escala foi a apresentada em Shin, Si e Dhillon (2012). Neste trabalho, os autores propõe uma aproximação multiescala do grafo para obter múltiplas visões granulares da rede. Para realizar a predição de *links* de uma forma escalável e precisa a partir de combinações em múltiplas escalas, os autores desenvolveram o *Multi-Scale Link Prediction* (MSLP). O trabalho combinou medidas de proximidade para realizar a predição de múltipla escala usando agrupamento hierárquico. O experimento com três conjuntos de dados reais demonstrou a eficácia do método apresentado, sendo que a combinação de MSLP e Katz (MSLP-Katz) obteve o melhor desempenho em todos os três conjuntos de dados com melhorias significativas em relação ao Katz.

Uma pesquisa diferente das anteriores há o trabalho de Kunegis, Preusse e Schwagereit (2013), o qual busca prever *links* negativos (como adversário ou desconfiança) em uma rede social, usando apenas os *links* positivos (como amizade e confiança). Utilizam dois conjuntos de dados e métodos de predição de *links* baseados em centralidade e em proximidade. Os *links* negativos têm um pequeno valor agregado, mas são mensuráveis nas redes sociais que foram estudadas pelos autores. Neste trabalho, foram realizados dois experimentos: com *links* positivos conhecidos e *links* positivos e negativos conhecidos. Os resultados experimentais apresentados pelos autores mostram que os melhores resultados foram obtidos quando as ligações negativas são conhecidas, apesar de a diferença entre os métodos ser pequena.

Steurer e Trattner (2013) estudam a predição de relacionamentos de interações e de reciprocidade entre usuários nas redes sociais. Para isso, utilizaram características obtidas de redes sociais (características topológicas e características homofílicas) e informação de localização do usuário. Para os experimentos, foram utilizados dados obtidos do jogo

Second Life. Utilizaram o algoritmo de Regressão Logística binomial para classificação e os experimentos foram validados usando validação cruzada 10-*fold*. Segundo os autores, os recursos de dados de localização são uma grande fonte para predizer as interações entre usuários em redes sociais *online*, superando os dados da rede social significativamente. Entretanto, para predizer a reciprocidade, os dados da rede social se mostraram mais úteis do que os dados de localização. A principal conclusão deste trabalho é que os resultados de ambos os experimentos mostram que a previsibilidade das interações e reciprocidade entre os usuários da rede social do Second Life pode ser significativamente melhorada se o classificador for treinado em ambos os conjuntos de características.

Com um enfoque diferenciado, Leroy, Cambazoglu e Bonchi (2010) pesquisaram a predição de *links* em um ambiente *cold start*, isto é, em um ambiente no qual se tem ainda poucas informações as relações entre as entidades (por exemplo, uma entidade que acabou de ser inserida no ambiente e ainda não possui nenhuma ligação com outras). Neste ambiente, os autores buscam predizer possíveis ligações através da exploração de outros tipos de informações disponíveis. Para tal, é proposto um método de duas fases com base em *bootstrap probabilistic graph* (BPG), na qual a primeira fase prevê a existência de um *link* e a segunda aplica as medidas baseadas no grafo para a predição final. Para testar esse método, foram utilizados conjuntos de dados do Flickr e o resultado obtido foi comparado com outros métodos considerados tradicionais. Segundo os autores, os experimentos demonstram a eficácia do método proposto para a predição de *links* neste ambiente. Eles concluem que, quanto mais informação, melhor é a precisão da predição.

Hsieh et al. (2013) também trabalharam em ambiente *cold start* e *warm start*. Os autores se referem ao *cold start* como uma situação em que o usuário acabou de entrar na rede. O *warm start* é um momento específico da rede, no qual a conexão entre os usuários é conhecida. Os autores deste trabalho propuseram um modelo matemático de afinidade entre usuários, calculando a probabilidade da conexão entre dois nós com base na sobreposição organizacional de uma empresa. O modelo foi validado experimentalmente com base em dados reais de uma rede social e pode ser utilizado tanto para prever *links* quanto para detectar comunidades/grupos. Segundo os autores, o desempenho da solução proposta é melhor do que CN e AA para prever *links* em ambos os ambientes, demonstrando que a relevância da sobreposição do tempo e da estrutura organizacional é importante na análise de predição.

3.3 Considerações finais

Com a revisão realizada é visível que o tema "Predição de relacionamentos em redes sociais" ainda é recente e suas bases teóricas ainda estão sendo firmadas. Isso é claro ao observar que a maioria das publicações apresenta e testa novas técnicas e atributos para uma melhor eficiência e/ou eficácia na predição. Mas, ao analisar todos como um conjunto, pode-se inferir que há alguns atributos (ou métricas) considerados tradicionais como CN, Katz, JC, AA e PA, pois são utilizados como base para comparar o desempenho das propostas em cada publicação. Esses atributos medem a similaridade ou a proximidade entre os nós relativos à topologia da rede.

Pode-se verificar que o tema predição de *links* é amplo, podendo ser aplicado nas mais diversas áreas onde há interação entre entidades, como humana, sistemas tecnológicos e sistemas biológicos (DONG et al., 2012; LIN; YUN; ZHU, 2012; RODRIGUEZ; ROGATI, 2012) e não apenas em redes sociais. O sistema de estabelecimento de relacionamentos é muito complexo, além da dinamicidade, às vezes engloba áreas diferentes, demonstrando a interdisciplinaridade do assunto. Em Quercia e Capra (2009), por exemplo, ao analisarem seus resultados de geolocalização, os autores perceberam a necessidade de considerar outros fatores como interesses, ocupação, preferências culturais, dentre outros, para uma análise completa do desenvolvimento de uma rede de relacionamento social.

Existem fatores externos à informação contida na rede social estudada que podem ter relação na dinâmica dos relacionamentos. Existe uma linha tênue entre os relacionamentos em uma rede *online* e os eventos *off-line*, nos quais o dia-a-dia da pessoa tem influência no comportamento da rede (RODRIGUEZ; ROGATI, 2012). E, também, as interações de usuário entre as diferentes redes *online*, onde o comportamento dos usuários nas mais diferentes redes sociais a qual pertence, pode complementar a fonte de informação para a tarefa de predição de *links*. Nessa linha de pesquisa são utilizadas as redes heterogêneas.

Além dos fatores externos, existe um fator intrínseco na maioria das redes sociais, devido ao dinamismo de crescimento em relação ao tempo. Com isso, o estudo fica mais complexo. Para lidar com isso, alguns trabalhos incluem em seus métodos a influência temporal, como em Gao, Denoyer e Gallinari (2011).

No que se refere às propostas de novos atributos ou métodos para predição de *links*, existem alguns desafios para se propor um método que obtenha bons resultados. Contudo, o bom desempenho do método em uma rede social específica não garante que o mesmo

terá um bom desempenho em outros contextos. Portanto, propor um método que seja flexível, em termos de aplicação, é um desafio. Além disso, os cálculos envolvidos em alguns métodos possuem alta complexidade computacional, tornando inviável sua aplicação em casos com enormes quantidades de dados, situação encontrada em muitas das redes sociais reais (SONG et al., 2012). Consequentemente, seria ideal que os métodos também fossem escaláveis. Esses desafios são perceptíveis na maioria dos trabalhos estudados durante a revisão sistemática, os quais tentam abordar pelo menos um desses desafios ou comentam a necessidade de se abordar em trabalhos futuros.

Por este levantamento não levar em conta apenas as redes acadêmicas (criadas com dados do arXiv, DBLP, PubMed, etc.), foi possível se obter uma visão mais abrangente, já que somente 33% dos artigos encontrados (16 artigos) abordam esse contexto. É provável que os atributos propostos em trabalhos que utilizaram outro tipo de rede social possam ser aplicados no contexto de relacionamentos acadêmicos (como coautoria).

Após a leitura dos artigos, algumas palavras poderiam ser sugeridas para futuras buscas tais como: *social graph*, *suggesting friends*, *recommendation friends* e *online media*.

Deste modo, pode-se concluir que há muitas possibilidades a serem pesquisadas sobre o assunto, ainda mais porque o tema é recente e está ganhando destaque nos últimos anos. Por isso, surgem muitas novas pesquisas com propostas de atributos, aplicações e de estudos para entendimento do relacionamento em uma rede social, sendo tal situação observada no presente trabalho.

4 Metodologia

Com base no estudo de trabalhos correlatos foram estabelecidos os atributos a serem utilizados bem como as ferramentas e parâmetros iniciais para a filtragem dos dados e a predição propriamente dita. Com base nestas informações e no trabalho de Digiampietri, Santiago e Alves (2013), as seguintes atividades foram consideradas necessárias para a realização dos testes da solução proposta: seleção da amostra, obtenção e armazenamento dos dados, identificação das informações relevantes, seleção dos atributos, filtragem dos dados, montagem dos conjuntos de treinamento e teste, execução dos testes e análise dos resultados.

4.1 Revisão da literatura e identificação das técnicas e atributos utilizados

Esta atividade foi realizada utilizando-se a metodologia de revisão sistemática, conforme apresentado no Capítulo 3. Foram identificados diversos atributos/métricas especialmente atributos estruturais calculados a partir da topologia de redes sociais (tais como: Adamic-Adar, Vizinhos em Comum, Conexão Preferencial, entre outros). Também foram identificados alguns atributos específicos do domínio das redes sociais acadêmicas potencialmente úteis para a predição de coautorias (por exemplo, relação de orientação; existência de orientados em comum; existência de orientadores em comum; atuação no mesmo programa de pós-graduação; áreas de interesse em comum).

4.2 Atividades realizadas nos experimentos

Esta subseção apresenta as atividades que foram realizadas durante a execução dos testes e validações.

4.2.1 Seleção da amostra

Este projeto está contextualizado dentro do Grupo de Análise de Redes Sociais e Cientometria (GARSC)[1] o qual tem como um de seus objetivos analisar os dados de todos

[1] http://plsql1.cnpq.br/buscaoperacional/detalhegrupo.jsp?grupo=0067103NX4DPZ6

os currículos Lattes[2] disponíveis. Este grupo possui um banco de dados com cerca de 4,2 milhões de currículos.

Para a amostra foram utilizados os dados disponibilizados publicamente na Plataforma Lattes. Foram selecionados 657 pesquisadores permanentes dos programas de pós-graduação em Ciência da Computação com doutorado e/ou mestrado acadêmico que foram avaliados nos triênios 2004 a 2006 e 2007 a 2009 pela CAPES (Coordenação e Aperfeiçoamento de Pessoal de Nível Superior).

A seleção desta amostra foi motivada pela presença de diferentes relacionamentos entre os pesquisadores como relações de orientação, orientados em comum, orientadores em comum, atuação no mesmo programa de pós-graduação, áreas de interesse em comum, relações de coautoria, dentre outros.

4.2.2 Obtenção e armazenamento dos dados

Ao longo do desenvolvimento deste projeto, duas metodologias diferentes foram utilizadas para a obtenção e organização inicial dos dados. Na primeira, os currículos são baixados da Internet no formato HTML diretamente da Plataforma Lattes do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), utilizando-se a ferramenta *wget*. O conteúdo de cada um dos arquivos HTML é convertido para XML, com a utilização de *parsers*, e o conjunto de arquivos XML é utilizado para criar um banco de dados relacional. Para esse processo são utilizadas as ferramentas desenvolvidas em Digiampietri et al. (2012a) e Digiampietri et al. (2012b). No banco de dados, as produções bibliográficas são discriminadas por tipo de publicação (artigo completo publicado em anais, artigos publicados em periódicos, etc.) e também as orientações estão organizadas por tipo (doutorado, mestrado, iniciação científica, etc.).

A segunda abordagem consiste em utilizar cópias dos arquivos obtidos da Plataforma Lattes no formato XML. Periodicamente o GARSC baixa cópias atualizadas destes arquivos. Os atributos apresentados no próximo capítulo foram calculados utilizando dados obtidos com a segunda abordagem.

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

4.2.3 Identificação das informações relevantes

As informações consideradas relevantes que foram selecionadas são: identificação de todos os currículos relacionados a cada pesquisador da amostra (incluindo coautores, orientados, orientadores, coparticipantes em bancas e em projetos de pesquisa), identificação dos orientandos e orientadores, identificação dos artigos publicados e identificação das coautorias. Estas informações foram utilizadas para a determinação dos atributos/características utilizados como entrada dos algoritmos de seleção de atributos e classificação a fim de se realizar a predição de relacionamentos e identificação das características mais importantes para esta predição. Para a identificação dos currículos relacionados a cada pesquisador, foram utilizadas as relações explícitas existentes em cada currículo, isto é, os *links* HTML existentes na Plataforma Lattes para indicar o currículo de um coautor, coparticipante de um projeto, coparticipante de uma banca, orientador ou orientando.

4.2.4 Seleção dos atributos

Neste trabalho foram consideradas apenas duas classes: "serão coautores" e "não serão coautores". Para verificar se dois artigos presentes em diferentes currículos correspondem a uma única publicação empregou-se a metodologia de resolução de entidades proposta em Digiampietri et al. (2012b), a qual especifica o tratamento de publicações cadastradas na Plataforma Lattes. Para verificar se dois pesquisadores possuem uma relação de orientação ou para saber quantos (co)orientadores em comum os dois pesquisadores possuem, aplicou-se o algoritmo de normalização de nomes presentes em registros bibliográficos proposto por Mugnaini et al. (2012). Foram implementados os algoritmos para calcular os seguintes atributos (métricas ou características): CN, SA, JC, AA, RA, SO, HPI, HDI, LHN, PA, KATZ e SP. A descrição de todos os atributos utilizados está presente na Tabela 5, enquanto a definição matemática dos atributos estruturais pode ser encontrada na subseção 2.3.1.

4.2.5 Filtragem horizontal de dados

A montagem do conjunto de treinamento envolve combinar os pesquisadores dois a dois e extrair os atributos selecionados para cada par de pesquisadores. Porém, um

grande volume de dados poderá ser produzido, impossibilitando o tratamento pela maioria dos classificadores. Para diminuir o volume, foi realizada uma filtragem horizontal dos dados, excluindo alguns pares de pesquisadores antes do processo de treinamento. Dois critérios de filtragem de dados foram testados, verificando-se o quão bom cada critério foi em termos de separar realmente apenas pares que não serão coautores e também em termos de redução do volume de dados original. Destaca-se que muitas das métricas só consideram como elementos candidatos à predição de relacionamentos aqueles que estejam em um mesmo componente conexo do grafo.

4.2.6 Montagem dos conjuntos de treinamento e de teste

Para o conjunto de treinamento os anos de 1971 a 2000 foi considerado passado, de 2001 a 2005 foi considerado presente e de 2006 a 2010 foi considerado futuro. No conjunto de teste, os anos de 1976 a 2005 foi considerado passado, de 2006 a 2010 foi considerado presente e de 2011 a 2015 foi considerado futuro. A janela definida como futuro determina os rótulos dos vetores de características, ou seja, as coautorias que devem ser preditas. A Figura 8 ilustra as janelas de tempo utilizadas para montar o conjunto de treinamento e de teste.

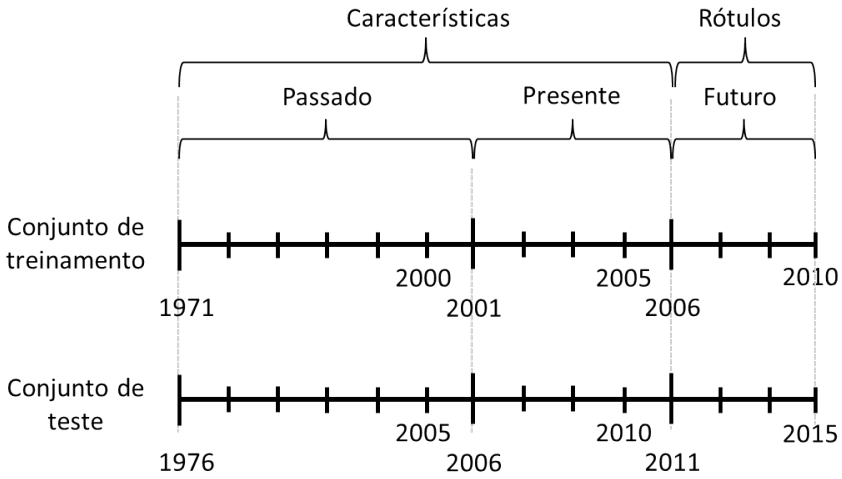
Diferentes conjuntos de treinamento foram montados de acordo com aplicação de técnicas de pré-processamento. Uma técnica aplicada foi a filtragem vertical dos dados, a qual excluiu um ou mais atributos dos pares selecionados. A montagem desses conjuntos obteve-se com a execução de seletores de atributos. A execução de testes utilizando os diferentes conjuntos formados permitiu a identificação dos atributos mais relevantes para a predição, bem como indicou se algum atributo utilizado não contribuirá com esta atividade. Ademais, para lidar com o grande desbalanceamento das classes, foram montados conjuntos de treinamento balanceados com a técnica de *Oversampling*.

Os atributos AA e RA são calculados em conjuntos que não são disjuntos (isto é, $\Gamma(x) \cap \Gamma(y) \neq \emptyset$). Portanto, o valor do atributo é zero se os conjuntos forem disjuntos.

4.2.7 Execução dos testes

Os testes foram realizados utilizando versões já implementadas de algoritmos de classificação disponíveis no Weka (HALL et al., 2009), que contém várias implementações de

Figura 8 – Representação das janelas de tempo para criação dos conjuntos de treinamento de teste.



Fonte: William T. Maruyama, 2015.

algoritmos para a seleção de atributos, classificação, agrupamento e identificação de regras de associação. Foram executados testes com os algoritmos de classificação disponíveis que classificam conjuntos de dados cuja classe é não numérica.

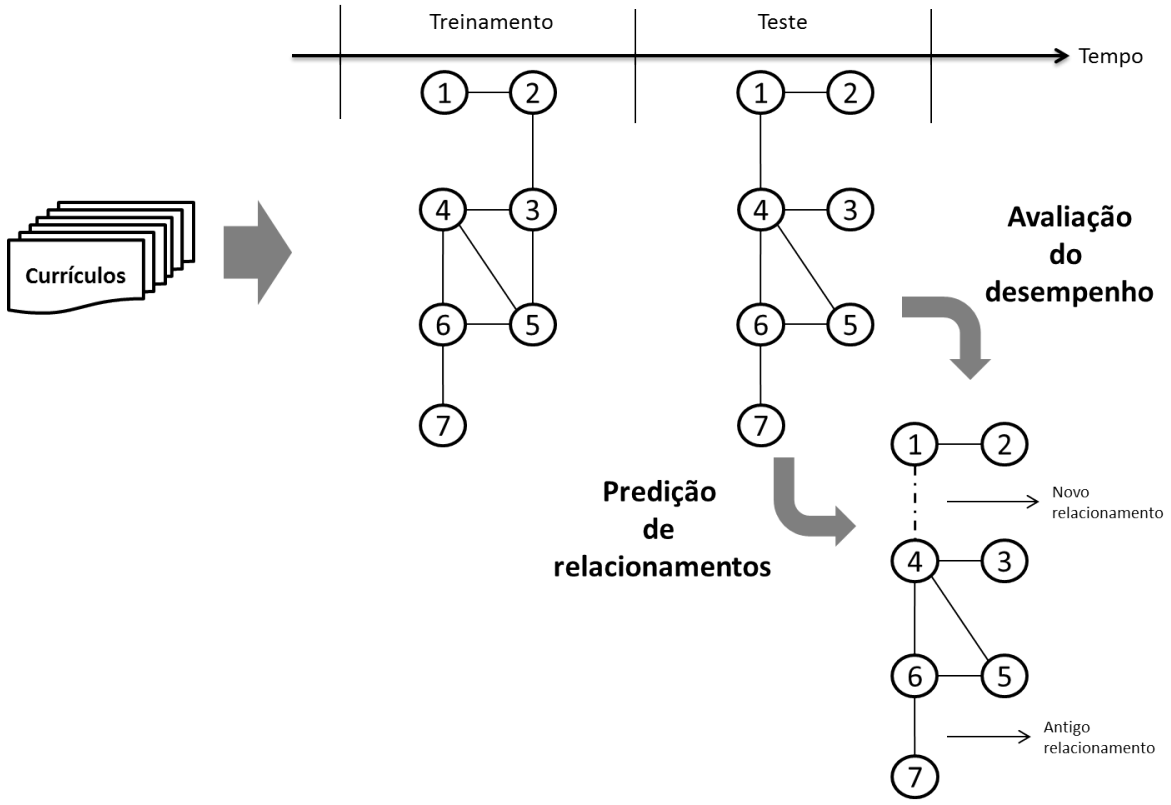
Foram realizados testes de predição no caso geral, ou seja, independentemente de serem inéditas ou reincidentes. Além disso, foram executados testes para verificar a predição de coautorias entre dois pesquisadores que não colaboraram na publicação de artigos dentro da janela de tempo do treinamento (relacionamentos inéditos).

A Figura 9 ilustra o processo de predição de coautorias utilizado. Com as informações obtidas dos currículos dos pesquisadores, os atributos dos conjuntos de teste e de treinamento são extraídos ou calculados a partir da rede de coautoria. No conjunto de treinamento, deste exemplo tem-se 8 relacionamento (arestas) sendo que calcula-se 21 possíveis relacionamentos (combinação dos 7 autores tomados aos pares). Com o conjunto de treinamento os algoritmos de classificação são treinados e o conjunto de teste utilizado para a predição, no qual o desempenho é medido e avaliado.

4.2.8 Solução desenvolvida

Com base nas atividades apresentadas, foi desenvolvida uma solução para os experimentos de predição de relacionamentos de coautorias, na linguagem Java. Com um arquivo de configuração é possível determinar o fluxo de execução e os parâmetros das funções a serem executadas. Com essa configuração é possível executar funções isoladas,

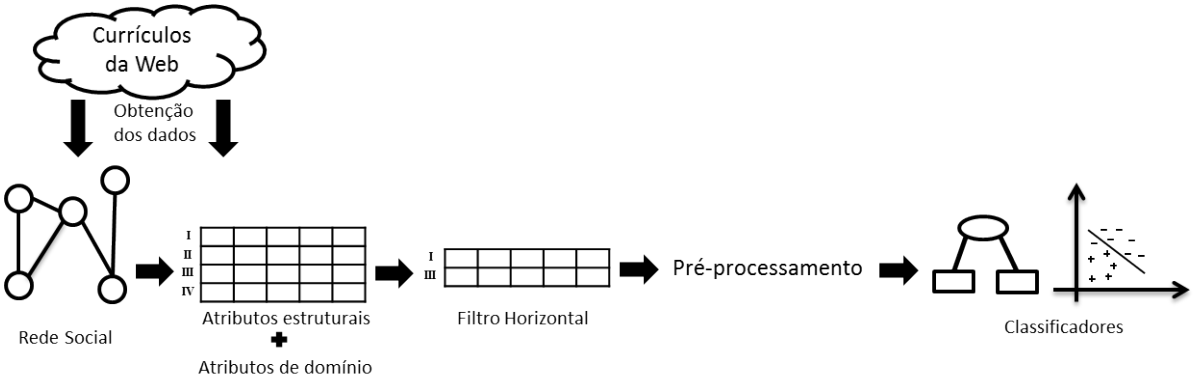
Figura 9 – Ilustração da predição de coautorias.



Fonte: William T. Maruyama, 2015.

ou todas as funções em sequência. A Figura 10 ilustra o processo de predição de coautorias que é utilizado neste trabalho.

Figura 10 – Processo de predição de coautorias da solução desenvolvida.



Fonte: William T. Maruyama, 2015.

Deste modo, o sistema permite a execução automatizada do processo de predição de coautorias. O sistema possui as seguintes funções:

1. Criar o conjunto de atributos: o sistema extrai e calcula alguns atributos de domínio dos dados, obtidos do Currículo Lattes, e também calcula atributos estruturais do grafo criado de relacionamentos de coautorias;
2. Normalização dos dados: por configuração é possível escolher entre dois tipos de normalização dos dados. Normalização por mínimo e máximo e *Z-Score*;
3. Balanceamento de dados: na configuração é possível realizar o balanceamento de dados utilizando a técnica de *Oversampling*;
4. Seleção de atributos: foi realizada a integração com a API do Weka. Enquanto um arquivo de configuração desta função determina os algoritmos e opções desejadas;
5. Classificação: foi realizada a integração com a API do Weka. Enquanto um arquivo de configuração desta função determina os algoritmos e opções desejadas.

Para todas as funções são criados arquivos de saída com o resultado obtido da execução.

Tabela 5 – Descrição dos atributos.

Tipo	Nome do atributo	Descrição
Estrutural	AA	Adamic-Adar - índice que atribui peso na relação de duas pessoas favorecendo as relações entre pessoas que possuem poucos relacionamentos (o peso do relacionamento é calculado pela somatória de 1 divido pelo logaritmo do número de relacionamentos [grau] dos vizinhos em comum destas duas pessoas).
	CN passado e presente	Quantidade de vizinhos em comum entre os pesquisadores na rede de coautorias formada por produções bibliográficas do passado e presente.
	CN presente	Quantidade de vizinhos em comum entre os pesquisadores na rede de coautorias formada por produções bibliográficas do presente.
	HDI	Hub Depressed Index - índice calculado pela divisão do número de elementos da intersecção de dois conjuntos dividido pelo número máximo de elementos entre estes dois conjuntos (por exemplo, número de vizinhos em comum de duas pessoas dividido pelo número máximo de vizinhos destas pessoas).
	HPI	Hub Promoted Index - índice calculado pela divisão do número de elementos da intersecção de dois conjuntos dividido pelo número mínimo de elementos entre estes dois conjuntos (por exemplo, número de vizinhos em comum de duas pessoas dividido pelo número mínimo de vizinhos destas pessoas).
	JC	Jaccard's Coefficient - índice que mede a similaridade entre dois conjuntos dividindo o número de elementos da intersecção dos dois conjuntos pelo número de elementos das união (por exemplo, número de vizinhos em comum dividido pela união dos vizinhos de duas pessoas).
	KATZ_0,0005 KATZ_0,005 KATZ_0,05	Katz é um índice calculado de maneira iterativa para estimar a influência de um par de nós em uma rede considerando-se os caminhos existentes entre os nós. Para este cálculo existe a necessidade da definição de uma constante Beta. Neste artigo três valores de Beta foram considerados: 0,05; 0,005; e 0,0005.
	LHN	Leicht-Holme-Newman Index - índice calculado pelo número de elementos da intersecção de dois conjuntos dividido pela multiplicação do número de elementos de cada conjunto (por exemplo, número de vizinhos em comum dividido pela multiplicação do número de vizinhos de duas pessoas).
	PA	Preferential Attachment - índice dado pela multiplicação entre o número de elementos de dois conjuntos (por exemplo, multiplicação do número de vizinhos de duas pessoas).
	RA	Resource Allocation - índice que atribui peso na relação de duas pessoas favorecendo as relações entre pessoas que possuem poucos relacionamentos (o peso do relacionamento é calculado pela somatória de 1 divido pelo número de relacionamentos [grau] dos vizinhos em comum destas duas pessoas).
	SA	Salton Index - índice que mede a coocorrência de dois elementos dividido pela raiz quadrada da multiplicação da ocorrência de cada elemento. Em redes sociais pode ser usado para medir relação entre o número de vizinhos que duas pessoas têm em comum divido pela raiz quadradada da multiplicação do número de vizinhos de cada um.
	SO	Sørensen Index - índice calculado como sendo duas vezes a intersecção entre dois conjuntos dividido pela soma dos elementos de cada conjunto (por exemplo, número de vizinhos em comum dividido pelo número de vizinhos da primeira pessoa mais o número de vizinhos da segunda).
	SP (Distância no grafo)	Shortest Path - caminho mínimo entre dois nós da rede.
Domínio/Contexto	Artigos em anais 1	Quantidade de artigos completos publicados em anais de conferências no período presente pela pessoa 1.
	Artigos em anais 2	Quantidade de artigos completos publicados em anais de conferências no período presente pela pessoa 2.
	Artigos em periódicos 1	Quantidade de artigos publicados em periódicos no período presente pela pessoa 1.
	Artigos em periódicos 2	Quantidade de artigos publicados em periódicos no período presente pela pessoa 2.
	Conferências passado	Quantidade de artigos publicados em conferências em coautorias pelo par de pesquisadores no passado.
	Conferências presente	Quantidade de artigos publicados em conferências em coautorias pelo par de pesquisadores no presente.
	Distância geográfica	Distância geográfica entre os endereços profissionais de dois pequisadores.
	Orientação em andamento	Atributo que recebe o valor 1 caso um dos pesquisadores seja orientador, em uma orientação em andamento, ou 0 caso contrário.
	Orientação passado	Atributo que recebe valor 1 caso um dos pesquisadores tenham sido orientador do outro no passado, ou 0 caso contrário.
	Orientação presente	Atributo que recebe valor 1 caso um dos pesquisadores tenham sido orientador do outro no presente, ou 0 caso contrário.
	Orientadores em comum	Quantidade de orientadores e coorientadores que foram orientadores dos dois pesquisadores em análise.
	Periódicos passado	Quantidade de artigos publicados em periódicos em coautorias pelo par de pesquisadores no passado.
	Periódicos presente	Quantidade de artigos publicados em periódicos em coautorias pelo par de pesquisadores no presente.
	Programas em comum	Atributo que recebe o valor 1 caso os dois pesquisadores pertençam ao mesmo programa de pós-graduação, ou 0 caso contrário.
	Subáreas em comum	Número de subáreas de atuação que os dois pesquisadores possuem em comum.

Fonte: William T. Maruyama, 2015.

5 Resultados e Discussão

Neste capítulo são apresentados os resultados dos experimentos e a discussão sobre os mesmos.

5.1 Resultados dos experimentos

Os experimentos testaram dois problemas de predição: o problema geral e o de novas coautorias. O primeiro analisa todos os possíveis *links*, independente se os autores já colaboraram ou não anteriormente. O segundo problema refere-se às novas coautorias, isto é, as coautorias inéditas na rede.

Para ambos os problemas, foram testadas duas abordagens: (I) o conjunto de treinamento possui apenas as instâncias resultantes do filtro horizontal, (II) ao conjunto de treinamento da abordagem I foram adicionadas as instâncias positivas (serão coautores) que haviam sido eliminadas pelo filtro. O filtro utilizado excluiu todas as instâncias (pares de pesquisadores) cuja maioria dos atributos tivesse valor nulo.

Quanto ao conjunto de dados, primeiramente foi analisado o conjunto de todos os atributos, sem o uso de nenhuma técnica de seleção de atributos ou redução de dimensionalidade. Após, foram realizados testes com variações na montagem dos (sub)conjuntos de atributos (conjunto estrutural, conjunto de domínio, seleção de atributos e com os atributos individualmente). Além disso, foram realizados testes sem e com balanceamento, utilizando a técnica *Oversampling* no conjunto de treinamento.

Como se trata de um problema de classificação binária tem-se duas possíveis classes: a classe "serão coautores" - ou classe positiva representada nas tabelas por T - e "não serão coautores" - ou classe negativa representada como F. Todos os resultados foram ranqueados decrescentemente conforme os valores de acurácia, revocação da classe T, AUC (*Area Under the Curve*) e Medida-F da classe T.

5.2 Problema geral

A seguir são apresentadas os resultados do problema geral. Esta seção foi subdividida em:

- 5.2.1 Abordagem I:

5.2.1 Abordagem I

A quantidade de instâncias por classe do conjunto de dados da abordagem I no problema geral de predição, com o conjunto de treinamento não balanceado, é apresentada na Tabela 6. No conjunto de treinamento balanceado, o número de instâncias da classe minoritária (classe T) é igualado à quantidade da classe majoritária (classe F).

Caso todas as instâncias fossem classificadas como pertencentes à classe negativa, a acurácia (valor base para as análises comparativas) seria de 95,24%.

Tabela 6 – Quantidade de instâncias da abordagem I no problema geral.

	Classe	
	F	T
Conjunto de treinamento	10955	878
Conjunto de teste	14425	721

Fonte: William T. Maruyama, 2015.

5.2.1.1 Abordagem I com todos os atributos no problema geral

A seguir, são apresentados os algoritmos que obtiveram os melhores valores de acurácia (Tabela 7), de revocação da classe T (Tabela 8), de AUC (Tabela 9) e de Medida-F (Tabela 10), no teste com todos os atributos - conjunto completo - sem balanceamento do conjunto de treinamento.

O algoritmo que obteve melhor acurácia no caso geral foi o *Attribute Selected Classifier*, sendo sua taxa de acerto de 96,091% no conjunto de teste (Tabela 7), apresentando

Tabela 7 – Três melhores resultados de acurácia com todos os atributos da abordagem I no problema geral.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
meta.Attribute SelectedClassifier	F	0,778	0,98	0,971	0,988	0,587	96,091
	T	0,778	0,502	0,638	0,413	0,012	
	Avg	0,778	0,957	0,955	0,961	0,559	
trees.ADTree	F	0,865	0,98	0,965	0,995	0,73	96,065
	T	0,865	0,396	0,736	0,27	0,005	
	Avg	0,865	0,952	0,954	0,961	0,695	
trees.BFTree	F	0,782	0,98	0,964	0,996	0,742	96,045
	T	0,782	0,383	0,744	0,258	0,004	
	Avg	0,782	0,951	0,954	0,96	0,707	

Fonte: William T. Maruyama, 2015.

melhor desempenho do que se todas as instâncias fossem classificadas como da classe negativa (95,24%). O desempenho da revocação deste algoritmo também foi o melhor dentre os três primeiros ranqueados em relação às identificações dos casos positivos. Ele identificou mais de 40% dos casos positivos e, das instâncias classificadas como positivas, o algoritmo acertou mais de 60% das vezes (Tabela 7). Contudo, apesar do 2º e 3º algoritmos terem identificados menos, estes foram mais precisos sobre os casos classificados como positivos.

Tabela 8 – Três melhores resultados de revocação da classe positiva com todos os atributos da abordagem I no problema geral.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
misc.VFI	F	0,829	0,693	0,989	0,533	0,118	54,998
	T	0,829	0,157	0,086	0,882	0,467	
	Avg	0,829	0,668	0,946	0,55	0,134	
bayes.BayesNet	F	0,881	0,921	0,986	0,864	0,24	85,924
	T	0,881	0,34	0,219	0,76	0,136	
	Avg	0,881	0,894	0,95	0,859	0,235	
bayes.NaiveBayes Updateable	F	0,874	0,943	0,985	0,905	0,28	89,601
	T	0,87	0,397	0,274	0,72	0,095	
	Avg	0,873	0,917	0,951	0,896	0,271	

Fonte: William T. Maruyama, 2015.

O algoritmo *VFI* foi o algoritmo de melhor desempenho em revocação (0,882), porém sua acurácia (54,998%) e precisão (0,086) foram relativamente baixos quando comparados ao 2º e 3º ranqueados (Tabela 8). O valor da revocação da classe negativa do *VFI* também apresentou o menor desempenho (0,533) dentre os três.

Em relação à AUC, o algoritmo *DMNBtext* obteve o melhor resultado do teste (0,886, Tabela 9). Sua acurácia de 95,444% apresentou valor maior que o valor base do problema geral (95,24%), contudo foi menor do que registrado pelo *Logit Boost*. O *DMNBtext* identificou 0,455 dos casos positivos (revocação) e atingiu 0,525 de acertos dos casos positivos (precisão).

Tabela 9 – Três melhores resultados de AUC com todos os atributos da abordagem I no problema geral.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
bayes.DMNBtext	F	0,886	0,976	0,973	0,979	0,545	95,444
	T	0,886	0,487	0,525	0,455	0,021	
	Avg	0,886	0,953	0,952	0,954	0,52	
bayes.BayesNet	F	0,881	0,921	0,986	0,864	0,24	85,924
	T	0,881	0,34	0,219	0,76	0,136	
	Avg	0,881	0,894	0,95	0,859	0,235	
meta.LogitBoost	F	0,88	0,979	0,971	0,986	0,589	95,88
	T	0,88	0,487	0,598	0,411	0,014	
	Avg	0,88	0,955	0,953	0,959	0,562	

Fonte: William T. Maruyama, 2015.

Tabela 10 – Três melhores resultados da Medida-F com todos os atributos da abordagem I no problema geral.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
meta.ThresholdSelector	F	0,865	0,972	0,979	0,966	0,413	94,758
	T	0,865	0,516	0,46	0,587	0,034	
	Avg	0,865	0,951	0,954	0,948	0,395	
meta.Classification ViaClustering	F	0,79	0,97	0,981	0,959	0,379	94,322
	T	0,79	0,51	0,433	0,621	0,041	
	Avg	0,79	0,948	0,955	0,943	0,363	
meta.Classification ViaRegression	F	0,875	0,979	0,972	0,986	0,567	95,999
	T	0,875	0,507	0,613	0,433	0,014	
	Avg	0,875	0,957	0,955	0,96	0,541	

Fonte: William T. Maruyama, 2015.

O maior valor da Medida-F foi 0,516, obtido com o algoritmo *Threshold Selector*, no qual os valores de precisão e revocação ficaram próximos a 0,5 (Tabela 10).

5.2.1.2 Abordagem I com todos atributos e balanceamento no problema geral

Nesta subseção são apresentados os resultados dos testes com todos os atributos e com balanceamento do conjunto de treinamento. Estes resultados foram ranqueados decrescentemente e os melhores três valores de acurácia (Tabela 11), revocação da classe positiva (Tabela 12), AUC (Tabela 13) e Medida-F (Tabela 14) são apresentados.

Tabela 11 – Três melhores resultados de acurácia com todos atributos da abordagem I no problema geral, com balanceamento.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
meta.RandomCommittee	F	0,822	0,977	0,97	0,984	0,605	95,629
	T	0,822	0,463	0,558	0,395	0,016	
	Avg	0,822	0,953	0,951	0,956	0,577	
meta.RotationForest	F	0,829	0,975	0,973	0,977	0,534	95,286
	T	0,829	0,485	0,505	0,466	0,023	
	Avg	0,829	0,952	0,951	0,953	0,51	
rules.ZeroR	F	0,5	0,976	0,952	1	1	95,24
	T	0,5	0	0	0	0	
	Avg	0,5	0,929	0,907	0,952	0,952	

Fonte: William T. Maruyama, 2015.

Ao se realizar o balanceamento no conjunto de treinamento, o algoritmo *Random Committee* foi o que apresentou melhor acurácia (95,629%, Tabela 11). Este valor foi maior do que o valor base (95,24%), mas inferior à acurácia do teste do conjunto que não foi balanceado (96,091%, Tabela 7). Contudo houve um aumento do valor de AUC entre o conjunto não balanceado e o balanceado (de 0,778 para 0,822, respectivamente).

Tabela 12 – Três melhores resultados de revocação da classe positiva com todos atributos da abordagem I no problema geral, com balanceamento.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
meta.StackingC	F	0,5	0	0	0	0	4,76
	T	0,5	0,091	0,048	1	1	
	Avg	0,5	0,004	0,002	0,048	0,048	
bayes.DMNBtext	F	0,744	0	0	0	0	4,76
	T	0,744	0,091	0,048	1	1	
	Avg	0,744	0,004	0,002	0,048	0,048	
meta.Classification ViaClustering	F	0,518	0,105	0,983	0,056	0,019	9,97
	T	0,518	0,094	0,049	0,981	0,944	
	Avg	0,518	0,105	0,938	0,1	0,063	

Fonte: William T. Maruyama, 2015.

Diferentemente da acurácia, o balanceamento do conjunto de treinamento apresentou melhora para a métrica de revocação (valor 1, Tabela 12), quando comparado com o teste de todos os atributos com conjunto de dados não balanceados (0,882, Tabela 8). Entretanto, apesar do elevado valor de revocação no caso positivo, a precisão do conjunto balanceado apresentou valores muito baixos (por exemplo, 0,048 para o algoritmo *StackingC*, o qual classificou todas as instâncias como positivas).

Tabela 13 – Três melhores resultados de AUC com todos atributos da abordagem I no problema geral, com balanceamento.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
trees.LADTree	F	0,884	0,943	0,985	0,905	0,273	89,641
	T	0,884	0,4	0,276	0,727	0,095	
	Avg	0,884	0,917	0,951	0,896	0,265	
trees.ADTree	F	0,88	0,936	0,986	0,891	0,259	88,34
	T	0,88	0,377	0,253	0,741	0,109	
	Avg	0,88	0,909	0,951	0,883	0,252	
bayes.BayesNet	F	0,88	0,9	0,987	0,826	0,215	82,424
	T	0,88	0,298	0,184	0,785	0,174	
	Avg	0,88	0,871	0,949	0,824	0,213	

Fonte: William T. Maruyama, 2015.

Os maiores valores de AUC no conjunto balanceado e no não balanceado ficaram próximos, sendo o primeiro colocado no ranqueamento do conjunto balanceado o menor entre eles (0,884 e 0,886 respectivamente, Tabelas 13 e 9). Porém, a revocação da classe positiva foi superior (0,727 contra 0,455).

Tabela 14 – Três melhores resultados de Medida-F com todos atributos da abordagem I no problema geral, com balanceamento.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
lazy.LWL	F	0,862	0,968	0,981	0,956	0,365	94,038
	T	0,862	0,504	0,417	0,635	0,044	
	Avg	0,862	0,946	0,954	0,94	0,35	
meta.MultiBoostAB	F	0,847	0,968	0,981	0,956	0,365	94,038
	T	0,847	0,504	0,417	0,635	0,044	
	Avg	0,847	0,946	0,954	0,94	0,35	
trees.DecisionStump	F	0,795	0,968	0,981	0,956	0,365	94,038
	T	0,795	0,504	0,417	0,635	0,044	
	Avg	0,795	0,946	0,954	0,94	0,35	

Fonte: William T. Maruyama, 2015.

O maior valor obtido da Medida-F com balanceamento foi 0,504 com o algoritmo *LWL* (Tabela 14). No geral, os valores atingidos ficaram próximos dos obtidos sem balanceamento (Tabela 10), mas com leve vantagem no teste sem balanceamento.

5.2.1.3 Abordagem I com atributos de domínio no problema geral

Nesse experimento foi testado o desempenho com o conjunto de atributos de domínio conforme apresentado na Tabela 5.

A seguir, são apresentados os resultados ranqueados de acurácia (Tabela 15), de revocação (Tabela 16), de AUC (Tabela 17) e de Medida-F (Tabela 18) do conjunto de atributos de domínio.

Tabela 15 – Três melhores resultados de acurácia dos atributos de domínio da abordagem I no problema geral.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
meta.FilteredClassifier	F	0,796	0,98	0,97	0,991	0,616	96,19
	T	0,796	0,49	0,676	0,384	0,009	
	Avg	0,796	0,957	0,956	0,962	0,587	
functions.Logistic	F	0,836	0,98	0,968	0,992	0,659	96,144
	T	0,836	0,457	0,693	0,341	0,008	
	Avg	0,836	0,955	0,955	0,961	0,628	
functions.Simple Logistic	F	0,836	0,98	0,968	0,992	0,66	96,131
	T	0,836	0,455	0,69	0,34	0,008	
	Avg	0,836	0,955	0,955	0,961	0,629	

Fonte: William T. Maruyama, 2015.

A maior acurácia foi registrado pelo algoritmo *Filtered Classifier* (96,19%), que é maior do que o valor base (95,24%). Aliás, os três melhores valores são maiores do que a acurácia base (Tabela 15). O conjunto de atributos de domínio apresentou valores de acurácia e de revocação superiores ao do conjunto completo de atributos sem balanceamento (Tabelas 15 e 7, respectivamente). Contudo, os valores de precisão da classe positiva deste

teste foram menores do que os registrados no conjunto com todos os atributos sem balanceamento.

Tabela 16 – Três melhores resultados da revocação da classe positiva dos atributos de domínio da abordagem I no problema geral.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
misc.VFI	F	0,859	0,818	0,988	0,698	0,171	70,467
	T	0,859	0,211	0,121	0,829	0,302	
	Avg	0,859	0,789	0,947	0,705	0,177	
trees.DecisionStump	F	0,775	0,974	0,979	0,97	0,42	95,114
	T	0,775	0,53	0,489	0,58	0,03	
	Avg	0,775	0,953	0,955	0,951	0,402	
rules.ConjunctiveRule	F	0,775	0,974	0,979	0,97	0,42	95,114
	T	0,775	0,53	0,489	0,58	0,03	
	Avg	0,775	0,953	0,955	0,951	0,402	

Fonte: William T. Maruyama, 2015.

Diferentemente, os maiores valores de revocação deste teste são menores do que os observados na análise com todos os atributos sem balanceamento (Tabelas 16 e 8, respectivamente). Contudo, a precisão e a acurácia deste ranqueamento foram maiores com relação àqueles apresentados na Tabela 8.

Tabela 17 – Três melhores resultados de AUC dos atributos de domínio da abordagem I no problema geral.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
bayes.BayesNet	F	0,86	0,971	0,975	0,967	0,492	94,54
	T	0,86	0,47	0,437	0,508	0,033	
	Avg	0,86	0,947	0,95	0,945	0,47	
misc.VFI	F	0,859	0,818	0,988	0,698	0,171	70,467
	T	0,859	0,211	0,121	0,829	0,302	
	Avg	0,859	0,789	0,947	0,705	0,177	
bayes.DMNBtext	F	0,859	0,978	0,971	0,984	0,587	95,728
	T	0,859	0,479	0,571	0,413	0,016	
	Avg	0,859	0,954	0,952	0,957	0,559	

Fonte: William T. Maruyama, 2015.

Do mesmo modo que os três melhores valores de revocação, os valores de AUC também são menores quando comparados com os resultados do teste com todos os atributos e sem balanceamento (Tabelas 17 e 9, respectivamente). Nesta linha de comparação, os valores registrados de precisão e acurácia são menores, mas os de revocação, são maiores com os atributos de domínio.

Os valores obtidos na Medida-F foram próximos àqueles do teste com todos os atributos (Tabela 10), mas com uma pequena vantagem para os resultados obtidos com o conjunto de atributos de domínio (Tabela 18).

Tabela 18 – Três melhores resultados da Medida-F dos atributos de domínio da abordagem I no problema geral.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
lazy.LWL	F	0,823	0,974	0,979	0,97	0,42	95,114
	T	0,823	0,53	0,489	0,58	0,03	
	Avg	0,823	0,953	0,955	0,951	0,402	
trees.DecisionStump	F	0,775	0,974	0,979	0,97	0,42	95,114
	T	0,775	0,53	0,489	0,58	0,03	
	Avg	0,775	0,953	0,955	0,951	0,402	
rules.ConjunctiveRule	F	0,775	0,974	0,979	0,97	0,42	95,114
	T	0,775	0,53	0,489	0,58	0,03	
	Avg	0,775	0,953	0,955	0,951	0,402	

Fonte: William T. Maruyama, 2015.

5.2.1.4 Abordagem I com atributos estruturais no problema geral

Esta subseção apresenta os resultados dos testes realizados a partir do conjunto de atributos estruturais (descrição na Tabela 5). Como nas seções anteriores, a seguir pode ser observado o ranqueamento conforme os valores de acurácia (Tabela 19), de revocação da classe positiva (Tabela 20), de AUC (Tabela 21) e de Medida-F (Tabela 22).

Tabela 19 – Três melhores resultados de acurácia do conjunto com atributos estruturais da abordagem I no problema geral.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
functions.Multilayer Perceptron	F	0,845	0,976	0,961	0,99	0,795	95,266
	T	0,845	0,292	0,507	0,205	0,01	
	Avg	0,845	0,943	0,94	0,953	0,757	
rules.ZeroR	F	0,5	0,976	0,952	1	1	95,24
	T	0,5	0	0	0	0	
	Avg	0,5	0,929	0,907	0,952	0,952	
rules.Ridor	F	0,528	0,976	0,955	0,997	0,942	95,24
	T	0,528	0,104	0,5	0,058	0,003	
	Avg	0,528	0,934	0,933	0,952	0,897	

Fonte: William T. Maruyama, 2015.

O algoritmo *Multilayer Perceptron* foi o que obteve maior acurácia (95,266%), sendo ela levemente superior ao valor base (95,24%). Contudo, o 2º e o 3º ranqueados apresentaram o mesmo valor que a acurácia base, isso devido as suas respectivas revocações da classe negativa serem altas. Neste caso, pode-se destacar o algoritmo *ZeroR* que classificou todas as instâncias como negativas, obtendo 95,24% de acurácia e 0,952 de precisão na classe negativa (Tabela 19).

No ranqueamento da revocação, o algoritmo *VFI* apresentou a melhor revocação da classe positiva, entretanto, com as menores acurácia e precisão dentre os três resultados (Tabela 20).

Ao observar comparativamente estes resultados com o obtido com todos os atributos e sem o balanceamento, a maior revocação de ambas as análises foram iguais (0,882,

Tabela 20 – Três melhores resultados de revocação da classe positiva do conjunto com atributos estruturais da abordagem I no problema geral.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
misc.VFI	F	0,809	0,635	0,988	0,468	0,118	48,759
	T	0,809	0,141	0,077	0,882	0,532	
	Avg	0,809	0,611	0,944	0,488	0,138	
bayes.BayesNet	F	0,848	0,903	0,987	0,832	0,227	82,906
	T	0,848	0,301	0,187	0,773	0,168	
	Avg	0,848	0,874	0,948	0,829	0,225	
bayes.NaiveBayesSimple	F	0,838	0,937	0,985	0,894	0,276	88,571
	T	0,838	0,376	0,254	0,724	0,106	
	Avg	0,838	0,91	0,95	0,886	0,268	

Fonte: William T. Maruyama, 2015.

Tabelas 20 e 8, respectivamente). Com relação aos atributos de domínio, os valores apresentados pelos atributos estruturais foram maiores (Tabelas 16 e 20, respectivamente).

Tabela 21 – Três melhores resultados de AUC do conjunto com atributos estruturais da abordagem I no problema geral.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
meta.LogitBoost	F	0,874	0,975	0,964	0,986	0,743	95,141
	T	0,874	0,335	0,481	0,257	0,014	
	Avg	0,874	0,944	0,941	0,951	0,709	
trees.LADTree	F	0,87	0,975	0,966	0,984	0,695	95,18
	T	0,87	0,376	0,49	0,305	0,016	
	Avg	0,87	0,946	0,943	0,952	0,663	
meta.Bagging	F	0,87	0,973	0,968	0,978	0,644	94,857
	T	0,87	0,398	0,449	0,356	0,022	
	Avg	0,87	0,946	0,943	0,949	0,614	

Fonte: William T. Maruyama, 2015.

Já para os valores de AUC, os resultados foram maiores que os obtidos com atributos de domínio e inferiores aos obtidos com o conjunto completo de atributos (Tabelas 21, 17 e 9, respectivamente).

Tabela 22 – Três melhores resultados da Medida-F do conjunto com atributos estruturais da abordagem I no problema geral.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
meta.Dagging	F	0,791	0,97	0,981	0,959	0,379	94,322
	T	0,791	0,51	0,433	0,621	0,041	
	Avg	0,791	0,948	0,955	0,943	0,363	
meta.Classification ViaClustering	F	0,79	0,97	0,981	0,959	0,379	94,322
	T	0,79	0,51	0,433	0,621	0,041	
	Avg	0,79	0,948	0,955	0,943	0,363	
functions.SMO	F	0,79	0,97	0,981	0,959	0,379	94,322
	T	0,79	0,51	0,433	0,621	0,041	
	Avg	0,79	0,948	0,955	0,943	0,363	

Fonte: William T. Maruyama, 2015.

Os melhores resultados da Medida-F no presente teste foram um pouco superiores aos valores do primeiro teste (Tabelas 22 e 10), mas ficaram próximos. Em relação ao teste com o conjunto de domínio, os valores foram inferiores ao conjunto estrutural (Tabelas 18 e 22).

De modo geral, os resultados com o conjunto de atributos estruturais obtiveram acurácia inferior aos atributos de domínio e ao conjunto completo. Contudo, verificou-se que a revocação foi superior aos atributos de domínio e próximos dos obtidos com o conjunto completo.

5.2.1.5 Abordagem I com seleção de atributos no problema geral

Por meio do uso de alguns algoritmos de seleção de atributos, foram encontrados os atributos mais relevantes para a predição de coautorias. A Tabela 23 apresenta cada subconjunto formado - conforme os respectivos algoritmos de seleção do arcabouço Weka e o método de busca utilizado que retornaram subconjuntos não vazios - e os atributos selecionados que os compõem (representados por "x").

Por meio da metodologia empregada, foram formados seis subconjuntos de atributos. Os Subconjunto 3 e 4 são os que apresentam maior quantidade de atributos selecionados (24 atributos) e o Subconjunto 5 o que apresenta a menor quantidade (3 atributos). Em alguns casos (Subconjuntos 1, 3 e 5), um mesmo subconjunto foi formado ao se utilizar um mesmo algoritmo de seleção, mas com diferentes métodos de busca. Os atributos "Periódicos presente" e "Conferências presente" estavam presentes em todos os seis subconjuntos e, contrariamente, "Artigos em periódicos 1", "CN presente" e "AA" não foram selecionados em nenhum dos subconjuntos formados.

Os resultados a seguir são apresentados segundo o ranqueamento do maior valor de acurácia (Tabela 24), de revocação (Tabela 25), de AUC (Tabela 26) e de Medida-F (Tabela 27), conforme análise de cada subconjunto de atributos selecionados, sem balanceamento do conjunto de treinamento. Isto é, são apresentados os três subconjuntos com melhores desempenho em cada uma das métricas avaliadas.

O melhor desempenho registrado no raqueamento da acurácia é do Subconjunto 1, ao se utilizar o classificador *DTNB* (96,243%, Tabela 24). Levando-se em consideração que o valor base é de 95,24%, a acurácia dos ranqueados foram maiores do que se todas as instâncias fossem consideradas negativas. Observa-se que o melhor resultado de acurácia obtido com a seleção de atributos foi superior ao resultado obtido utilizando-se todos os atributos sem balanceamento (o melhor valor havia sido 96,091%, Tabela 7).

A melhor revocação da classe positiva registrada foi a do Subconjunto 3, contudo a sua precisão foi baixa e, consequentemente, com um baixo valor de acurácia (0,963, 0,55 e

Tabela 23 – Subconjuntos obtidos com os algoritmos de seleção de características da abordagem I no problema geral.

	Subconjuntos					
Atributos	1	2	3	4	5	6
Periódicos passado			x			
Conferências passado			x	x		
Periódicos presente	x	x	x	x	x	x
Conferências presente	x	x	x	x	x	x
Orientação passado			x	x		
Orientação presente			x			
Orientação em andamento		x	x			x
Orientadores em comum			x	x		
Orientandos em comum			x	x		
CN passado e presente		x	x	x		x
Programas em comum			x	x		
Artigos em periódicos 1						
Artigos em anais 1			x	x		
Artigos em periódicos 2			x	x		
Artigos em anais 2			x	x		
CN presente						
SA			x	x		
JC			x	x		
AA						
RA			x	x		
SO				x		
HPI			x	x		
HDI			x	x		
LHN				x		
PA			x	x		
KATZ 0.05			x			
KATZ 0.005		x	x	x		x
KATZ 0.0005				x		
Subáreas em comum			x	x		
Distância geográfica	x		x	x		
Distância no grafo (SP)	x	x		x	x	x

Nota: Os números dos subcojuntos correspondem aos algoritmos testados. O primeiro algoritmo é o de seleção seguido do(s) método(s) de busca utilizado(s):
(1) *CfsSubsetEval* / *GreedyStepwise*, *BestFirst* e *LinearForwardSelection*
(2) *CfsSubsetEval* / *GeneticSearch*
(3) *ConsistencySubsetEval* / *GreedyStepwise*, *BestFirst* e *LinearForwardSelection*
(4) *ConsistencySubsetEval* / *GeneticSearch*
(5) *FilteredSubsetEval* / *GreedyStepwise*, *BestFirst* e *LinearForwardSelection*
(6) *FilteredSubsetEval* / *GeneticSearch*

Fonte: William T. Maruyama, 2015.

Tabela 24 – Os melhores resultados de acurácia em relação aos primeiros colocados em cada subconjunto de atributos da abordagem I no problema geral.

Subconjunto	Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
1	rules.DTNB	F	0,816	0,98	0,972	0,99	0,58	96,243
		T	0,816	0,516	0,667	0,42	0,01	
		Avg	0,816	0,958	0,957	0,962	0,553	
5	functions.Multilayer Perceptron	F	0,826	0,98	0,966	0,994	0,693	96,151
		T	0,826	0,431	0,727	0,307	0,006	
		Avg	0,826	0,954	0,955	0,962	0,661	
4	meta.Classification ViaRegression	F	0,879	0,98	0,971	0,989	0,591	96,111
		T	0,879	0,5	0,644	0,409	0,011	
		Avg	0,879	0,957	0,955	0,961	0,563	

Fonte: William T. Maruyama, 2015.

Tabela 25 – Os melhores resultados da revocação da classe positiva em relação aos primeiros colocados em cada subconjunto de atributos da abordagem I no problema geral.

Subconjunto	Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
3	misc.VFI	F	0,826	0,284	0,989	0,166	0,037	20,355
		T	0,826	0,103	0,055	0,963	0,834	
		Avg	0,826	0,275	0,944	0,204	0,075	
4	misc.VFI	F	0,839	0,762	0,989	0,62	0,133	63,192
		T	0,839	0,183	0,102	0,867	0,38	
		Avg	0,839	0,735	0,947	0,632	0,145	
6	misc.VFI	F	0,851	0,791	0,99	0,659	0,139	66,889
		T	0,851	0,198	0,112	0,861	0,341	
		Avg	0,851	0,763	0,948	0,669	0,148	

Fonte: William T. Maruyama, 2015.

20,355%, respectivamente, Tabela 25). Situação parecida foi observada no teste com todos os atributos sem balanceamento, no qual o valor da revocação foi menor que o apresentado no presente teste, mas a precisão e acurácia atuais foram maiores (Tabela 8).

Tabela 26 – Os melhores resultados de AUC em relação aos primeiros colocados em cada subconjunto de atributos da abordagem I no problema geral.

Subconjunto	Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
4	bayes.DMNBtext	F	0,884	0,977	0,974	0,979	0,523	95,543
		T	0,884	0,505	0,536	0,477	0,021	
		Avg	0,884	0,954	0,953	0,955	0,499	
3	bayes.DMNBtext	F	0,884	0,976	0,972	0,981	0,562	95,477
		T	0,884	0,48	0,53	0,438	0,019	
		Avg	0,884	0,953	0,951	0,955	0,536	
2	trees.LADTree	F	0,883	0,979	0,97	0,989	0,613	95,999
		T	0,883	0,479	0,63	0,387	0,011	
		Avg	0,883	0,955	0,954	0,96	0,584	

Fonte: William T. Maruyama, 2015.

O Subconjunto 4 foi o que apresentou melhor resultado quanto à AUC (0,884, Tabela 26), contudo este valor foi inferior ao registrado no teste com todos os atributos e sem balanceamento (0,886, Tabela 9). A despeito disso, o valor de revocação da classe positiva, a precisão e a acurácia deste subconjunto foram maiores quando comparados aos valores obtidos pela primeira colocação do ranqueamento de AUC no teste com o cojunto completo de atributos (Tabelas 26 e 9, respectivamente).

Tabela 27 – Os melhores resultados da Medida-F em relação aos primeiros colocados em cada subconjunto de atributos da abordagem I no problema geral.

Subconjunto	Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
1	lazy.LWL	F	0,825	0,974	0,979	0,97	0,42	95,114
		T	0,825	0,53	0,489	0,58	0,03	
		Avg	0,825	0,953	0,955	0,951	0,402	
3	meta.ThresholdSelector	F	0,863	0,972	0,98	0,963	0,393	94,632
		T	0,863	0,519	0,452	0,607	0,037	
		Avg	0,863	0,95	0,955	0,946	0,376	
2	meta.Attribute SelectedClassifier	F	0,778	0,979	0,973	0,986	0,555	96,058
		T	0,778	0,518	0,62	0,445	0,014	
		Avg	0,778	0,957	0,956	0,961	0,529	

Fonte: William T. Maruyama, 2015.

O maior resultado de Medida-F foi alcançado com o Subconjunto 1 (0,53), sendo este valor um pouco maior que o obtido com o conjunto total de atributos sem balanceamento (Tabelas 27 e 10, respectivamente).

De modo geral, pode-se considerar que o Subconjunto 4 foi o que apresentou melhor desempenho nas métricas analisadas, dentre todos os subconjuntos formados. Isto porque ele pode ser observado em 3º, 2º e 1º posições nos ranqueamentos de acurácia, revocação e de AUC (Tabelas 24, 25 e 26 respectivamente). Este subconjunto é formado por 12 atributos estruturais e 12 de domínio (Tabela 23 e 5) e foi selecionado a partir do algoritmo *Consistency Subset Eval* e do método de busca *Genetic Search*. Apesar de seu posicionamento e levando em consideração os ranqueamentos, somente o valor de acurácia foi maior do que os resultados obtidos no teste com todos os atributos sem balanceamento.

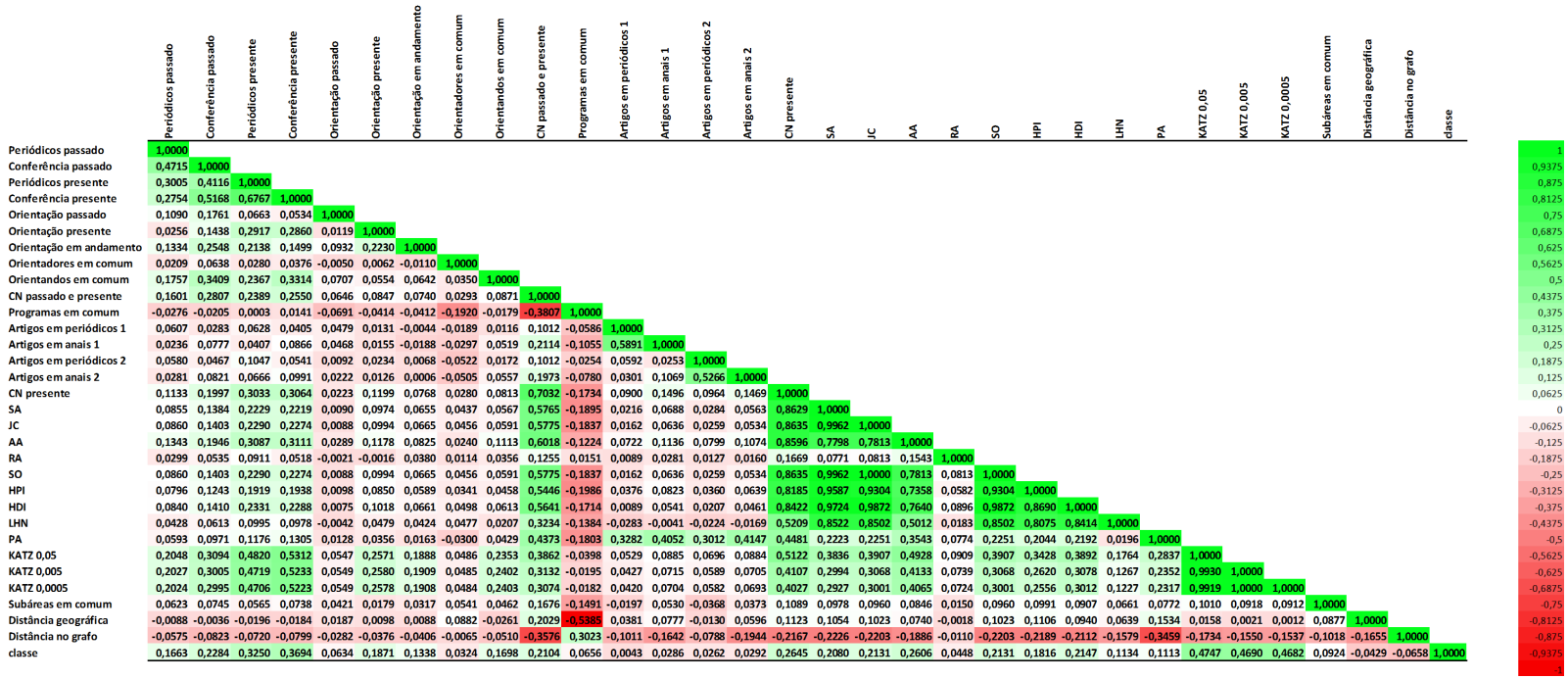
5.2.1.6 Abordagem I com atributos individuais no problema geral

A análise com os atributos individuais foi realizada de acordo com os diversos seletores de atributos presentes no arcabouço Weka, em relação à classe categórica cujos valores possíveis "não serão coautores" ou "serão coautores". Os algoritmos de seleção utilizam o critério de ranqueamento, que ordenam os atributos de acordo com sua importância. Isto é, a posição de cada atributo conforme a ordenação de cada algoritmo de seleção e apresentada na Tabela 28, onde 1 representa o atributo mais importante/informativo; 2, o segundo; e assim por diante. Na última coluna é apresentada a mediana da ordenação de cada atributo.

Conforme mediana, os atributos com melhor ordenação são "Conferências presente", "Katz" (as três variações) e "Distância no grafo". Os atributos com maiores valores de ordenação - portanto os menos informativos - foram "Artigos em periódicos" (as duas variações), "Artigos em anais" (ambas as variações) e "Orientadores em comum".

A Figura 11 contém a correlação entre a classe e os demais atributos. Para a identificação dos atributos mais relevantes será utilizada apenas a última linha da matriz de correlações (as demais linhas servem apenas para ilustrar algumas características do conjunto de dados), a qual apresenta a correlação de cada atributo com a *classe*. Para o cálculo das correlações, o valor do atributo classe "não serão coautores" foi substituído por 0 (zero) e o valor "serão coautores" foi substituído por 1 (um).

Figura 11 – Matriz de correlação dos atributos individuais no problema geral.



	Periódicos passado	Conferência passado	Periódicos presente	Conferência presente	Orientação passado	Orientação presente	Orientação em andamento	Orientadores em comum	Orientandos em comum	CN passado e presente	Programas em comum	Artigos em periódicos 1	Artigos em anais 1	Artigos em periódicos 2	Artigos em anais 2	CN presente	SA	JC	AA	RA	SO	HPI	HDI	LHN	PA	KATZ 0,05	KATZ 0,005	KATZ 0,0005	Subáreas em comum	Distância geográfica	Distância no grafo	classe
Periódicos passado	1,0000																															
Conferência passado	0,4715	1,0000																														
Periódicos presente	0,3005	0,4116	1,0000																													
Conferência presente	0,2754	0,5168	0,6767	1,0000																												
Orientação passado	0,1090	0,1761	0,0663	0,0534	1,0000																											
Orientação presente	0,0256	0,1438	0,2917	0,2860	0,0119	1,0000																										
Orientação em andamento	0,1334	0,2548	0,2138	0,1499	0,0932	0,2230	1,0000																									
Orientadores em comum	0,0209	0,0638	0,0280	0,0376	-0,0050	0,0062	-0,0110	1,0000																								
Orientandos em comum	0,1757	0,3409	0,2367	0,3314	0,0707	0,0554	0,0642	0,0350	1,0000																							
CN passado e presente	0,1601	0,2807	0,2389	0,2550	0,0646	0,0847	0,0740	0,0293	0,0871	1,0000																						
Programas em comum	-0,0276	-0,0205	0,0003	0,0141	-0,0691	-0,0414	-0,0412	-0,1920	-0,0179	-0,3807	1,0000																					
Artigos em periódicos 1	0,0607	0,0283	0,0628	0,0405	0,0479	0,0131	-0,0044	-0,0189	0,0116	0,1012	-0,0586	1,0000																				
Artigos em anais 1	0,0236	0,0777	0,0407	0,0866	0,0468	0,0155	-0,0188	-0,0297	0,0519	0,2114	-0,1055	0,5891	1,0000																			
Artigos em periódicos 2	0,0580	0,0467	0,1047	0,0541	0,0092	0,0234	0,0068	-0,0522	0,0172	0,1012	-0,0254	0,0592	0,0253	1,0000																		
Artigos em anais 2	0,0281	0,0821	0,0666	0,0991	0,0222	0,0126	0,0006	-0,0505	0,0557	0,1973	-0,0780	0,0301	0,1069	0,5266	1,0000																	
CN presente	0,1133	0,1997	0,3033	0,3064	0,0223	0,1199	0,0768	0,0280	0,0813	0,7032	-0,1734	0,0900	0,1496	0,0964	0,1469	1,0000																
SA	0,0855	0,1384	0,2229	0,2219	0,0090	0,0974	0,0655	0,0437	0,0567	0,5765	-0,1895	0,0216	0,0688	0,0284	0,0563	0,8629	1,0000															
JC	0,0860	0,1403	0,2290	0,2274	0,0088	0,0994	0,0665	0,0456	0,0591	0,5775	-0,1837	0,0162	0,0636	0,0259	0,0534	0,8635	0,9962	1,0000														
AA	0,1343	0,1946	0,3087	0,3111	0,0289	0,1178	0,0825	0,0240	0,1113	0,6018	-0,1224	0,0722	0,1136	0,0799	0,1074	0,8596	0,7798	0,7813	1,0000													
RA	0,0299	0,0535	0,0911	0,0518	-0,0021	-0,0016	0,0380	0,0114	0,0356	0,1255	0,0151	0,0089	0,0281	0,0127	0,0160	0,1669	0,0771	0,0813	0,1543	1,0000												
SO	0,0860	0,1403	0,2290	0,2274	0,0088	0,0994	0,0665	0,0456	0,0591	0,5775	-0,1837	0,0162	0,0636	0,0259	0,0534	0,8635	0,9962	1,0000	0,7813	0,0813	1,0000											
HPI	0,0796	0,1243	0,1919	0,1938	0,0098	0,0850	0,0589	0,0341	0,0458	0,5446	-0,1986	0,0376	0,0823	0,0360	0,0639	0,8185	0,9587	0,9304	0,7358	0,0582	0,9304	1,0000										
HDI	0,0840	0,1410	0,2331	0,2288	0,0075	0,1018	0,0661	0,0498	0,0613	0,5641	-0,1714	0,0089	0,0541	0,0207	0,0461	0,8422	0,9724	0,9872	0,7640	0,0896	0,9872	0,8690	1,0000									
LHN	0,0428	0,0613	0,0995	0,0978	-0,0042	0,0479	0,0424	0,0477	0,0207	0,3234	-0,1384	-0,0283	-0,0041	-0,0224	-0,0169	0,5209	0,8522	0,8502	0,5012	0,0183	0,8502	0,8075	0,8414	1,0000								
PA	0,0593	0,0971	0,1176	0,1305	0,0128	0,0356	0,0163	-0,0300	0,0429	0,4373	-0,1803	0,3282	0,4052	0,3012	0,4147	0,4481	0,2223	0,2251	0,3543	0,0774	0,2251	0,2044	0,2192	0,0196	1,0000							
KATZ 0,05	0,2048	0,3094	0,4820	0,5312	0,0547	0,2571	0,1888	0,0486	0,2353	0,3862	-0,0398	0,0529	0,0885	0,0696	0,0884	0,5122	0,3836	0,3907	0,4928	0,0909	0,3907	0,3428	0,3892	0,1764	0,2837	1,0000						
KATZ 0,005	0,2027	0,3005	0,4719	0,5233	0,0549	0,2580	0,1909	0,0485	0,2402	0,3132	-0,0195	0,0427	0,0715	0,0589	0,0705	0,4107	0,2994	0,3068	0,4133	0,0739	0,3068	0,2620	0,3078	0,1267	0,2352	0,9930	1,0000					
KATZ 0,0005	0,2024	0,2995	0,4706	0,5223	0,0549	0,2578	0,1908	0,0484	0,2403	0,3074	-0,0182	0,0420	0,0704	0,0582	0,0693	0,4027	0,2927	0,3001	0,4065	0,0724	0,3001	0,2556	0,3012	0,1227	0,2317	0,9919	1,0000	1,0000				
Subáreas em comum	0,0623	0,0745	0,0565	0,0738	0,0421	0,0179	0,0317	0,0541	0,0462	0,1676	-0,1491	-0,0197	0,0530	-0,0368	0,0373	0,1089	0,0978	0,0960	0,0846	0,0150	0,0960	0,0991	0,0907	0,0661	0,0772	0,1010	0,0918	0,0912	1,0000			
Distância geográfica	-0,0088	-0,0036	-0,0196	-0,0184	0,0187	0,0098	0,0088	0,0882	-0,0261	0,2029	-0,5385	0,0381	0,0777	-0,0130	0,0596	0,1123	0,1054	0,1023	0,0740	-0,0018	0,1023	0,1106	0,0940	0,0639	0,1534	0,0158	0,0021	0,0012	0,0877	1,0000		
Distância no grafo	-0,0575	-0,0823	-0,0720	-0,0799	-0,0282	-0,0376	-0,0406	-0,0065	-0,0510	-0,3576	0,3023	-0,1011	-0,1642	-0,0788	-0,1944	-0,2167	-0,2226	-0,2203	-0,1886	-0,0110	-0,2203	-0,2189	-0,2112	-0,1579	-0,3459	-0,1734	-0,1550	-0,1537	-0,1018	-0,1655	1,0000	
classe	0,1663	0,2284	0,3250	0,3694	0,0634	0,1871	0,1338	0,0324	0,1698	0,2104	0,0656	0,0043	0,0286	0,0262	0,0292	0,2645	0,2080	0,2131	0,2606	0,0448	0,2131	0,1816	0,2147	0,1134	0,1113	0,4747	0,4690	0,4682	0,0924	-0,0429	-0,0658	1,0000

Fonte: William T. Maruyama, 2015.

Tabela 28 – Ranqueamento dos atributos individuais da abordagem I no problema geral.

	ChiSquaredAttributeEval	*FilteredAttributeEval*	*GainRatioAttributeEval*	*InfoGainAttributeEval*	*OneRAttributeEval*	*ReliefFAttributeEval*	*SymmetricalUncertAttributeEval*	Mediana
Conferências presente	1	4	3	4	1	27	1	3
KATZ 0.05	2	1	9	1	3	24	6	3
KATZ 0.005	3	2	8	2	4	29	5	4
KATZ 0.0005	4	3	7	3	14	31	4	4
Distância no grafo (SP)	5	5	4	5	21	12	2	5
Periódicos presente	6	6	1	6	2	28	3	6
CN presente	8	9	14	9	7	25	11	9
AA	7	7	12	7	17	22	9	9
RA	10	8	13	8	18	20	10	10
HPI	9	10	16	10	5	11	14	10
CN passado e presente	11	11	11	11	13	9	8	11
HDI	12	12	20	12	8	13	15	12
Conferências passado	13	16	10	16	9	21	7	13
SA	14	13	21	13	10	14	19	14
SO	15	14	18	14	12	15	16	15
JC	16	15	19	15	16	16	17	16
Orientação presente	17	20	2	20	6	26	12	17
LHN	18	17	17	17	15	19	18	17
Periódicos passado	19	18	6	18	29	30	13	18
Orientandos em comum	20	19	15	19	11	17	20	19
Orientação em andamento	21	22	5	22	20	23	21	21
PA	22	21	23	21	31	10	22	22
Orientação passado	26	26	22	26	19	18	23	23
Subáreas em comum	23	23	25	23	24	7	25	23
Distância geográfica	24	24	24	24	25	2	24	24
Programas em comum	25	25	26	25	22	1	26	25
Orientadores em comum	27	30	27	30	30	8	27	27
Artigos em anais 2	29	27	28	27	26	3	28	27
Artigos em anais 1	28	28	29	28	28	4	29	28
Artigos em periódicos 2	30	29	30	29	27	5	30	29
Artigos em periódicos 1	31	31	31	31	23	6	31	31

Fonte: William T. Maruyama, 2015.

Os valores mais altos de correlação nesta linha são os atributos “Katz” (as três variações), “Conferências presente” e “Periódicos presente” (Figura 11). Destaca-se nesta tabela a grande correlação existente entre a maioria dos atributos derivados do grafo correspondente a rede social (atributos estruturais, Tabela 5).

De modo geral, pode-se notar que o atributo “Programas em comum” é o que apresenta maior número de correlações negativas com os demais atributos e sua correlação com o atributo “Distância geográfica” é a maior correlação negativa da matriz (Figura 11). Isso pode ser explicado pelo fato de, geralmente, a distância do endereço profissional ser baixa entre dois docentes que pertencem a um mesmo programa de pós-graduação e,

consequentemente, a distância ser maior quando os docentes não pertencem ao mesmo programa.

A seguir são apresentados os resultados da classificação ao utilizar cada atributo individualmente. Os resultados de cada atributo foram ranqueados para cada métrica analisada (acurácia, revocação da classe positiva, AUC e Medida-F) e são apresentados os três melhores atributos em relação a cada uma das métricas.

Tabela 29 – Três melhores atributos em relação à acurácia da abordagem I no problema geral.

Atributo	Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
Conferências presente	trees.NBTree	F	0,778	0,98	0,965	0,995	0,718	96,098
		T	0,778	0,407	0,736	0,282	0,005	
		Avg	0,778	0,953	0,954	0,961	0,684	
Periódicos presente	functions.Simple Logistic	F	0,67	0,978	0,961	0,997	0,81	95,827
		T	0,67	0,302	0,741	0,19	0,003	
		Avg	0,67	0,946	0,95	0,958	0,772	
Periódicos passado	trees.REPTree	F	0,602	0,977	0,956	0,998	0,914	95,484
		T	0,602	0,153	0,713	0,086	0,002	
		Avg	0,602	0,938	0,945	0,955	0,871	

Fonte: William T. Maruyama, 2015.

O melhor resultado em acurácia foi de 96,098%, obtido pelo atributo "Conferências presente" (Tabela 29). Este é um desempenho maior do que o valor base para a abordagem (95,24%) e é um pouco superior ao registrado pelo conjunto completo de atributos sem balanceamento (Tabela 7). Entretanto ele é inferior ao apresentado pelos subconjuntos selecionados de atributos (Tabela 24).

Levando-se em consideração que os três atributos listados na Tabela 29 são atributos de domínio (Tabela 5), a acurácia resultante do teste com o conjunto de atributos de domínio apresentou-se maior (Tabela 15) do que o registrado com a análise dos atributos individualmente.

Tabela 30 – Três melhores atributos em relação à revocação da classe positiva da abordagem I no problema geral.

Atributo	Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
Artigos em periódico 1	meta.Threshold Selector	F	0,562	0	0	0	0	4,76
		T	0,562	0,091	0,048	1	1	
		Avg	0,562	0,004	0,002	0,048	0,048	
Orientação passado	meta.Threshold Selector	F	0,528	0	0	0	0	4,76
		T	0,528	0,091	0,048	1	1	
		Avg	0,528	0,004	0,002	0,048	0,048	
Orientação em andamento	meta.Threshold Selector	F	0,524	0	0	0	0	4,76
		T	0,524	0,091	0,048	1	1	
		Avg	0,524	0,004	0,002	0,048	0,048	

Fonte: William T. Maruyama, 2015.

Os três melhores resultados em relação à revocação da classe positiva foram iguais quanto ao classificador (*Threshold Selector*), ao valor (1), à precisão (0,048) e, consequentemente, à acurácia (4,76%, Tabela 30).

Comparando tal resultado com outros testes realizados, este apresentou valores de revocação superiores aos registrados usando o conjunto completo de atributos e seleção de atributos (Tabela 30, 8 e 25, respectivamente), mas menores precisão e acurácia que os mesmos testes. Isso deve-se ao fato de que as intâncias foram todas classificadas como casos positivos.

Tabela 31 – Três melhores atributos em relação à AUC da abordagem I no problema geral.

Atributo	Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
KATZ 0,05	bayes.NaiveBayes Updateable	F	0,851	0,97	0,981	0,959	0,379	94,322
		T	0,851	0,51	0,433	0,621	0,041	
		Avg	0,851	0,948	0,955	0,943	0,363	
KATZ 0,005	lazy.LWL	F	0,85	0,975	0,96	0,992	0,832	95,24
		T	0,85	0,251	0,5	0,168	0,008	
		Avg	0,85	0,941	0,938	0,952	0,793	
KATZ 0,0005	lazy.KStar	F	0,844	0,976	0,952	1	1	95,24
		T	0,844	0	0	0	0	
		Avg	0,844	0,929	0,907	0,952	0,952	

Fonte: William T. Maruyama, 2015.

Em relação a AUC, os valores obtidos neste teste foram inferiores aos registrados com a utilização do conjunto completo e dos seletores de atributos (Tabela 31, 9 e 26 respectivamente). Considerando que os três atributos listados são estruturais (Tabela 5), este resultado também foi menor do que apresentado na análise dos atributos estruturais (Tabela 21).

Tabela 32 – Três melhores atributos em relação à Medida-F da abordagem I no problema geral.

Atributo	Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
Conferências presente	meta.Threshold Selector	F	0,778	0,974	0,979	0,97	0,42	95,114
		T	0,778	0,53	0,489	0,58	0,03	
		Avg	0,778	0,953	0,955	0,951	0,402	
KATZ 0,005	meta.Threshold Selector	F	0,85	0,97	0,981	0,959	0,376	94,322
		T	0,85	0,511	0,433	0,624	0,041	
		Avg	0,85	0,948	0,955	0,943	0,36	
KATZ 0,05	bayes.NaiveBayes Updateable	F	0,851	0,97	0,981	0,959	0,379	94,322
		T	0,851	0,51	0,433	0,621	0,041	
		Avg	0,851	0,948	0,955	0,943	0,363	

Fonte: William T. Maruyama, 2015.

O maior resultado na Medida-F foi alcaçado com o atributo “Conferências presente” (0,53). O valor foi igual ao obtido com o Subconjunto de atributos 1 (Tabela 27) e um pouco maior do que com o conjunto total de atributos sem balanceamento (Tabela 10).

5.2.1.7 Abordagem I com atributos individuais e balanceamento no problema geral

Do mesmo modo que a análise do conjunto completo de atributos, o conjunto de treinamento dos atributos individuais também foi balanceado com a técnica de *Oversampling* no presente teste.

A seguir são apresentados os resultados dos melhores atributos a partir do ranqueamento dos valores de acurácia (Tabela 33), de revocação da classe positiva (Tabela 34), de AUC (Tabela 35) e de Medida-F (Tabela 36).

Tabela 33 – Três melhores atributos em relação à acurácia da abordagem I no problema geral, com balanceamento.

Atributo	Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
Conferências presente	meta.Dagging	F	0,778	0,98	0,965	0,995	0,718	96,098
		T	0,778	0,407	0,736	0,282	0,005	
		Avg	0,778	0,953	0,954	0,961	0,684	
Periódicos presente	meta.Classification ViaClustering	F	0,593	0,978	0,961	0,997	0,81	95,827
		T	0,593	0,302	0,741	0,19	0,003	
		Avg	0,593	0,946	0,95	0,958	0,772	
Conferências passado	lazy.IB1	F	0,535	0,977	0,956	0,999	0,928	95,471
		T	0,535	0,132	0,754	0,072	0,001	
		Avg	0,535	0,937	0,946	0,955	0,884	

Fonte: William T. Maruyama, 2015.

O melhor desempenho da acurácia foi registrado pelo atributo "Conferências presente" (96,098%), seguido por outros dois atributos de domínio (Tabela 33). Os três resultados são superiores à acurácia base da abordagem I (95,24%). Estes valores e os melhores atributos são semelhantes aos observados na análise de atributos individuais não balanceado (Tabela 29).

Quanto à comparação com o teste realizado com todos os atributos e com balanceamento, os valores de acurácia dos atributos individuais foram superiores (Tabelas 11 e 33, respectivamente).

Tabela 34 – Três melhores atributos em relação à revocação da classe positiva da abordagem I no problema geral, com balanceamento.

Atributo	Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
KATZ 0,0005	bayes.DMNBtext	F	0,788	0	0	0	0	4,76
		T	0,788	0,091	0,048	1	1	
		Avg	0,788	0,004	0,002	0,048	0,048	
Conferências presente	bayes.DMNBtext	F	0,775	0	0	0	0	4,76
		T	0,775	0,091	0,048	1	1	
		Avg	0,775	0,004	0,002	0,048	0,048	
KATZ 0,005	bayes.DMNBtext	F	0,74	0	0	0	0	4,76
		T	0,74	0,091	0,048	1	1	
		Avg	0,74	0,004	0,002	0,048	0,048	

Fonte: William T. Maruyama, 2015.

Os valores de revocação da classe positiva são semelhantes aos obtidos pelo teste dos atributos individuais sem balanceamento do conjunto de dados, no qual todas as instâncias foram classificadas como positivas, o que diminuiu o valor da precisão e da acurácia (Tabelas 34 e 30, respectivamente). Contudo os atributos e o classificador listados são diferentes entre os dois resultados.

Tabela 35 – Três melhores atributos em relação à AUC da abordagem I no problema geral, com balanceamento.

Atributo	Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
KATZ 0,05	lazy.KStar	F	0,853	0,97	0,981	0,959	0,379	94,322
		T	0,853	0,51	0,433	0,621	0,041	
		Avg	0,853	0,948	0,955	0,943	0,363	
KATZ 0,005	lazy.KStar	F	0,85	0,97	0,981	0,959	0,379	94,322
		T	0,85	0,51	0,433	0,621	0,041	
		Avg	0,85	0,948	0,955	0,943	0,363	
KATZ 0,0005	trees.FT	F	0,844	0,97	0,981	0,959	0,379	94,322
		T	0,844	0,51	0,433	0,621	0,041	
		Avg	0,844	0,948	0,955	0,943	0,363	

Fonte: William T. Maruyama, 2015.

Como no caso da análise dos atributos individuais não balanceado, os mesmos três atributos estruturais registraram os melhores resultados - e na mesma ordem - de AUC (Tabelas 31 e 35, respectivamente). Contudo o valor de AUC do "Katz 0,05" foi um pouco superior no teste atual, assim como a revocação, precisão e consequente acurácia dos 2º e 3º colocados do ranqueamento também foram superiores.

Tabela 36 – Três melhores atributos em relação à Medida-F da abordagem I no problema geral, com balanceamento.

Atributo	Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
Conferências presente	trees.RandomTree	F	0,778	0,974	0,979	0,97	0,42	95,114
		T	0,778	0,53	0,489	0,58	0,03	
		Avg	0,778	0,953	0,955	0,951	0,402	
KATZ 0,05	bayes.NaiveBayes Updateable	F	0,851	0,97	0,981	0,959	0,376	94,322
		T	0,851	0,511	0,433	0,624	0,041	
		Avg	0,851	0,948	0,955	0,943	0,36	
KATZ 0,005	lazy.KStar	F	0,85	0,97	0,981	0,959	0,379	94,322
		T	0,85	0,51	0,433	0,621	0,041	
		Avg	0,85	0,948	0,955	0,943	0,363	

Fonte: William T. Maruyama, 2015.

O maior valor da Medida-F foi obtido com o atributo "Conferências presente" (0,53), sendo o atributo e o valor iguais ao teste sem balanceamento (Tabelas 36 e 32, respectivamente).

5.2.2 Abordagem II

Ainda considerando o problema geral de predição, a abordagem II refere-se à adição das instâncias positivas que foram eliminadas pelo filtro ao conjunto de instâncias resultante do filtro horizontal.

A quantidade de instâncias por classe do conjunto utilizada na abordagem II não balanceada pode ser observada na Tabela 37. A acurácia/valor base, isto é, a taxa de acerto geral de uma classificação toda negativa, nesta abordagem é de 95,24%.

Tabela 37 – Quantidade de instâncias da abordagem II no problema geral.

	Classe	
	F	T
Conjunto de treinamento	10955	1036
Conjunto de teste	14425	721

Fonte: William T. Maruyama, 2015.

5.2.2.1 Abordagem II com todos atributos no problema geral

A seguir, são apresentados os algoritmos que obtiveram a melhor taxa de acerto, conforme o ranqueamento da acurácia (Tabela 38), a revocação da classe positiva (Tabela 39), a AUC (Tabela 40) e a Medida-F (Tabela 41), do teste com todos os atributos e sem balanceamento do conjunto de dados.

Tabela 38 – Três melhores resultados de acurácia com todos atributos da abordagem II no problema geral.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
trees.ADTree	F	0,864	0,98	0,965	0,995	0,718	96,091
	T	0,864	0,407	0,733	0,282	0,005	
	Avg	0,864	0,953	0,954	0,961	0,684	
meta.Attribute SelectedClassifier	F	0,773	0,98	0,971	0,988	0,581	96,072
	T	0,773	0,504	0,632	0,419	0,012	
	Avg	0,773	0,957	0,955	0,961	0,554	
trees.LADTree	F	0,842	0,979	0,97	0,99	0,623	96,039
	T	0,842	0,476	0,643	0,377	0,01	
	Avg	0,842	0,955	0,954	0,96	0,594	

Fonte: William T. Maruyama, 2015.

A melhor acurácia foi registrada em 96,091% pelo classificador *ADTree* (Tabela 38), desempenho este, superior ao valor base da abordagem (95,24%). A revocação da classe positiva do 1º colocado foi o menor, quando comparado com os dois seguintes, contudo, sua precisão foi a maior.

Quanto à comparação com a abordagem I (Tabela 7), os resultados foram semelhantes quanto à 1ª posição, diferindo no valor de revocação e precisão - na abordagem I apresentou-se valor maior e menor, respectivamente.

Tabela 39 – Três melhores resultados de revocação da classe positiva com todos os atributos da abordagem II no problema geral.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
misc.VFI	F	0,825	0,724	0,989	0,571	0,132	58,497
	T	0,825	0,166	0,092	0,868	0,429	
	Avg	0,825	0,697	0,946	0,585	0,146	
bayes.NaiveBayes Updateable	F	0,869	0,942	0,985	0,903	0,279	89,423
	T	0,865	0,394	0,271	0,721	0,097	
	Avg	0,869	0,916	0,951	0,894	0,27	
bayes.NaiveBayes	F	0,869	0,942	0,985	0,903	0,279	89,423
	T	0,865	0,394	0,271	0,721	0,097	
	Avg	0,869	0,916	0,951	0,894	0,27	

Fonte: William T. Maruyama, 2015.

O classificador *VFI* foi o que obteve melhor resultado de revocação (0,868), mas dentre os listados, apresentou a menor precisão (0,092, Tabela 39). Contudo esta precisão foi maior, a despeito da sua revocação de valor inferior, quando comparado com a abordagem I (Tabela 8).

Tabela 40 – Três melhores resultados de AUC com todos os atributos da abordagem II no problema geral.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
bayes.DMNBtext	F	0,88	0,977	0,972	0,981	0,558	95,557
	T	0,88	0,487	0,541	0,442	0,019	
	Avg	0,88	0,953	0,952	0,956	0,532	
bayes.BayesNet	F	0,88	0,945	0,985	0,908	0,282	89,918
	T	0,88	0,404	0,281	0,718	0,092	
	Avg	0,88	0,919	0,951	0,899	0,273	
meta.Bagging	F	0,878	0,978	0,973	0,984	0,556	95,788
	T	0,878	0,501	0,575	0,444	0,016	
	Avg	0,878	0,955	0,954	0,958	0,53	

Fonte: William T. Maruyama, 2015.

O melhor desempenho de AUC registrado neste teste foi do classificador *DMNB text*, com 0,88 (Tabela 40). A abordagem atual apresentou valor de AUC menor do que o registrado na abordagem I (Tabela 9).

O algoritmo *Threshold Selector* foi o que apresentou maior valor de Medida-F (0,514, Tabela 41), sendo ele muito próximo ao teste do conjunto completo sem balanceamento da abordagem I - inclusive, o mesmo algoritmo foi o primeiro classificado (Tabela 10).

Tabela 41 – Três melhores resultados da Medida-F com todos os atributos da abordagem II no problema geral.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
meta.ThresholdSelector	F	0,828	0,971	0,98	0,962	0,39	94,513
	T	0,828	0,514	0,444	0,61	0,038	
	Avg	0,828	0,949	0,955	0,945	0,373	
meta.MultiBoostAB	F	0,786	0,969	0,981	0,956	0,366	94,104
	T	0,786	0,506	0,421	0,634	0,044	
	Avg	0,786	0,947	0,955	0,941	0,351	
trees.DecisionStump	F	0,795	0,968	0,981	0,956	0,365	94,038
	T	0,795	0,504	0,417	0,635	0,044	
	Avg	0,795	0,946	0,954	0,94	0,35	

Fonte: William T. Maruyama, 2015.

5.2.2.2 Abordagem II com todos atributos e balanceamento no problema geral

Do mesmo modo que a abordagem anterior, foi realizado o balanceamento do conjunto de treinamento utilizando a técnica *Oversampling*.

A seguir são apresentados os resultados obtidos neste respectivo teste, conforme o ranqueamento dos algoritmos que obtiveram a melhor taxa de acerto nas métricas: acurácia (Tabela 42), revocação (Tabela 43), AUC (Tabela 44) e Medida-F (Tabela 45).

Tabela 42 – Três melhores resultados de acurácia com todos os atributos da abordagem II no problema geral, com balanceamento.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
meta.RandomCommittee	F	0,827	0,976	0,969	0,984	0,632	95,471
	T	0,827	0,436	0,535	0,368	0,016	
	Avg	0,827	0,951	0,948	0,955	0,603	
meta.RotationForest	F	0,835	0,975	0,972	0,978	0,555	95,253
	T	0,835	0,472	0,502	0,445	0,022	
	Avg	0,835	0,951	0,95	0,953	0,529	
rules.ZeroR	F	0,5	0,976	0,952	1	1	95,24
	T	0,5	0	0	0	0	
	Avg	0,5	0,929	0,907	0,952	0,952	

Fonte: William T. Maruyama, 2015.

A maior acurácia foi de 95,471% do algoritmo *Random Committee* (Tabela 42), sendo ela - e a do 2º colocado - superior ao valor base do problema geral, de 95,24%. A revocação do 1º colocado foi menor que do 2º, com precisão um pouco acima de 0,5 (Tabela 42).

O processo de balanceamento pode ter influenciado negativamente nos resultados de acurácia, pois seus valores são menores quando comparado aos do não balanceado (Tabelas 42 e 38, respectivamente). Contudo, os valores de revocação foram um pouco maiores, excetuando-se os resultado do classificador *ZeroR*, que considerou todas as instâncias como negativas.

Quanto à comparação entre os dois tipos de abordagens, a melhor acurácia do teste atual foi um pouco menor com relação à análise do conjunto completo de atributos e com balanceamento da abordagem I (Tabela 42 e 11, respectivamente).

Tabela 43 – Três melhores resultados de revocação da classe positiva com todos os atributos da abordagem II no problema geral, com balanceamento.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
meta.StackingC	F	0,5	0	0	0	0	4,76
	T	0,5	0,091	0,048	1	1	
	Avg	0,5	0,004	0,002	0,048	0,048	
bayes.DMNBtext	F	0,67	0	0	0	0	4,76
	T	0,67	0,091	0,048	1	1	
	Avg	0,67	0,004	0,002	0,048	0,048	
meta.Classification ViaClustering	F	0,518	0,105	0,983	0,056	0,019	9,97
	T	0,518	0,094	0,049	0,981	0,944	
	Avg	0,518	0,105	0,938	0,1	0,063	

Fonte: William T. Maruyama, 2015.

O balanceamento favoreceu a melhora na revocação, pois classificou todas as instâncias como da classe positiva, como é o caso do 1º e 2º colocados, mas ambos possuem baixa precisão e, consequentemente, acurácia (Tabela 43).

Resultados semelhantes quanto a revocação, precisão, acurácia e algoritmos, foram registrados no teste de conjunto completo com balanceamento na abordagem I (Tabela 12).

Tabela 44 – Três melhores resultados de AUC com todos os atributos da abordagem II no problema geral, com balanceamento.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
bayes.BayesNet	F	0,876	0,909	0,987	0,843	0,229	83,983
	T	0,876	0,314	0,197	0,771	0,157	
	Avg	0,876	0,881	0,949	0,84	0,225	
trees.LADTree	F	0,872	0,92	0,986	0,862	0,24	85,745
	T	0,872	0,337	0,216	0,76	0,138	
	Avg	0,872	0,892	0,95	0,857	0,235	
bayes.NaiveBayesSimple	F	0,871	0,939	0,985	0,897	0,275	88,901
	T	0,866	0,384	0,261	0,725	0,103	
	Avg	0,871	0,913	0,95	0,889	0,266	

Fonte: William T. Maruyama, 2015.

Os valores de AUC foram superiores aos registrados na abordagem II sem balanceamento (Tabela 44 e 40, respectivamente). Porém foram inferiores aos resultados da abordagem I balanceada (Tabela 13).

O maior valor de Medida-F foi de 0,507, registrado pelo algoritmo *Dagging* (Tabela 45). Os valores obtidos da Medida-F foram inferiores aos obtidos sem balanceamento na abordagem II (Tabela 41), mas um pouco superiores aos obtidos pelo teste de conjunto completo e com balanceamento da abordagem I (Tabela 14).

Tabela 45 – Três melhores resultados da Medida-F com todos os atributos da abordagem II no problema geral, com balanceamento.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
meta.Dagging	F	0,795	0,969	0,981	0,958	0,374	94,203
	T	0,795	0,507	0,426	0,626	0,042	
	Avg	0,795	0,947	0,954	0,942	0,359	
functions.SMO	F	0,792	0,969	0,981	0,958	0,373	94,177
	T	0,792	0,506	0,424	0,627	0,042	
	Avg	0,792	0,947	0,954	0,942	0,357	
trees.DecisionStump	F	0,795	0,968	0,981	0,956	0,365	94,038
	T	0,795	0,504	0,417	0,635	0,044	
	Avg	0,795	0,946	0,954	0,94	0,35	

Fonte: William T. Maruyama, 2015.

5.3 Novas coautorias

O problema das novas coautorias é, no presente trabalho, considerado um segundo tipo de problema de predição. Este é focado somente na análise da formação de novos relacionamentos (ou *links* inéditos) na rede. Semelhante ao problema geral de predição, foram realizados diferentes testes/análises para uma melhor exploração sobre o assunto:

- 5.3.1 Abordagem I:
  - 5.3.1.1 Abordagem I com todos os atributos;
  - 5.3.1.2 Abordagem I com todos os atributos e balanceamento;
  - 5.3.1.3 Abordagem I com atributos de domínio;
  - 5.3.1.4 Abordagem I com atributos estruturais;
  - 5.3.1.5 Abordagem I com seleção de atributos;
  - 5.3.1.6 Abordagem I com atributos individuais;
  - 5.3.1.7 Abordagem I com atributos individuais e balanceamento.
- 5.3.2 Abordagem II:
  - 5.3.2.1 Abordagem II com todos atributos;
  - 5.3.2.2 Abordagem II com todos atributos e balanceamento.

5.3.1 Abordagem I

No problema de predição de novas coautorias, a quantidade de instâncias por classe do conjunto de dados utilizada na abordagem I (instâncias resultante do filtro horizontal) não balanceada pode ser observada na Tabela 46. Para o conjunto de treinamento balanceado, o número de instâncias no conjunto de treinamento da classe minoritária (classe T) é

igualado à quantidade da classe majoritária (classe F), utilizando a técnica *Oversampling*. Caso todas as instâncias fossem classificadas como negativas, a acurácia base (ou valor base) seria de 98,058%.

Tabela 46 – Quantidade de instâncias da abordagem I no problema de novas coautorias.

	Classe	
	F	T
Conjunto de treinamento	10537	439
Conjunto de teste	13838	274

Fonte: William T. Maruyama, 2015.

5.3.1.1 Abordagem I com todos atributos no problema de novas coautorias

A seguir são apresentados os algoritmos que obtiveram os melhores resultados conforme a acurácia (Tabela 47), a revocação da classe positiva (Tabela 48), a AUC (Tabela 49) e a Medida-F (Tabela 50) sem balanceamento do conjunto de treinamento.

Tabela 47 – Três melhores resultados de acurácia com todos os atributos da abordagem I no problema de novas coautorias.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
bayes.BayesianLogistic Regression	F	0,502	0,99	0,981	1	0,996	98,065
	T	0,502	0,007	1	0,004	0	
	Avg	0,502	0,971	0,981	0,981	0,977	
bayes.DMNBtext	F	0,751	0,99	0,981	1	0,985	98,058
	T	0,751	0,028	0,5	0,015	0	
	Avg	0,751	0,972	0,972	0,981	0,966	
trees.ADTree	F	0,725	0,99	0,981	1	1	98,058
	T	0,725	0	0	0	0	
	Avg	0,725	0,971	0,962	0,981	0,981	

Fonte: William T. Maruyama, 2015.

O melhor resultado de acurácia foi de 98,065%, valor um pouco maior do que o valor base (98,058%). No entanto, os valores de acurácia obtidas pelos algoritmos seguintes atingiram apenas 98,058%, pois não foram capazes de classificar casos positivos com precisão (Tabela 47).

O algoritmo *VFI* obteve o melhor valor de revocação (0,858), o qual foi o dobro do segundo colocado (Tabela 48). Contudo, os valores de verdadeiros positivos tiveram baixas taxas. Além disso, pode-se observar que, enquanto os valores de revocação diminuíram, os de precisão aumentaram.

Em relação à AUC, o algoritmo *Classification Via Regression* registrou o maior valor (0,752, Tabela 49). Além disso, os dois melhores colocados classificaram poucos casos

Tabela 48 – Três melhores resultados de revocação da classe positiva com todos os atributos da abordagem I no problema de novas coautorias.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
misc.VFI	F	0,71	0,537	0,992	0,368	0,142	37,791
	T	0,71	0,051	0,026	0,858	0,632	
	Avg	0,71	0,528	0,974	0,378	0,152	
meta.Classification ViaClustering	F	0,642	0,895	0,987	0,818	0,533	81,094
	T	0,642	0,088	0,048	0,467	0,182	
	Avg	0,642	0,879	0,969	0,811	0,526	
bayes.BayesNet	F	0,734	0,914	0,987	0,85	0,58	84,212
	T	0,734	0,094	0,053	0,42	0,15	
	Avg	0,734	0,898	0,969	0,842	0,572	

Fonte: William T. Maruyama, 2015.

Tabela 49 – Três melhores resultados de AUC com todos os atributos da abordagem I no problema de novas coautorias.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
meta.Classification ViaRegression	F	0,752	0,99	0,981	1	0,993	98,037
	T	0,752	0,014	0,286	0,007	0	
	Avg	0,752	0,971	0,967	0,98	0,973	
bayes.DMNBtext	F	0,751	0,99	0,981	1	0,985	98,058
	T	0,751	0,028	0,5	0,015	0	
	Avg	0,751	0,972	0,972	0,981	0,966	
bayes.NaiveBayes Updateable	F	0,742	0,949	0,986	0,914	0,675	90,271
	T	0,742	0,115	0,07	0,325	0,086	
	Avg	0,742	0,932	0,968	0,903	0,664	

Fonte: William T. Maruyama, 2015.

como positivos (menos de 0,01). Diferentemente, o algoritmo *Naive Bayes Updateable*, 3º, classificou mais casos como positivos, mas com baixa precisão (0,07, Tabela 49).

Tabela 50 – Três melhores resultados de Medida-F com todos os atributos da abordagem I no problema de novas coautorias.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
meta.ThresholdSelector	F	0,718	0,956	0,986	0,929	0,686	91,667
	T	0,718	0,128	0,08	0,314	0,071	
	Avg	0,718	0,94	0,968	0,917	0,674	
bayes.NaiveBayes Updateable	F	0,742	0,949	0,986	0,914	0,675	90,271
	T	0,742	0,115	0,07	0,325	0,086	
	Avg	0,742	0,932	0,968	0,903	0,664	
bayes.NaiveBayes	F	0,742	0,949	0,986	0,914	0,675	90,271
	T	0,742	0,115	0,07	0,325	0,086	
	Avg	0,742	0,932	0,968	0,903	0,664	

Fonte: William T. Maruyama, 2015.

O maior valor da Medida-F foi obtido pelo algoritmo *Threshold Selector* (0,128, Tabela 50). O valor baixo pode ser explicado pela revocação (0,314) e, principalmente, a precisão (0,08) serem baixas.

5.3.1.2 Abordagem I com todos atributos e balanceamento no problema de novas coautorias

Esta subseção apresenta os resultados obtidos pela predição de novas coautorias segundo a abordagem I e com o conjunto de treinamento balanceado. Os algoritmos foram ranqueados decrescentemente conforme a taxa de acerto (Tabela 51), a revocação (Tabela 52), a AUC (Tabela 53) e a Medida-F (Tabela 54).

Tabela 51 – Três melhores resultados de acurácia com todos os atributos na abordagem I no problema de novas coautorias, com balanceamento.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
rules.ZeroR	F	0,5	0,99	0,981	1	1	98,058
	T	0,5	0	0	0	0	
	Avg	0,5	0,971	0,962	0,981	0,981	
meta.Vote	F	0,5	0,99	0,981	1	1	98,058
	T	0,5	0	0	0	0	
	Avg	0,5	0,971	0,962	0,981	0,981	
meta.Stacking	F	0,5	0,99	0,981	1	1	98,058
	T	0,5	0	0	0	0	
	Avg	0,5	0,971	0,962	0,981	0,981	

Fonte: William T. Maruyama, 2015.

Ao realizar o balanceamento dos dados, não foi observada uma melhoria nos resultados de acurácia (Tabela 51) em relação ao não balanceado (Tabela 47). Nesta comparação, a acurácia do 1º colocado foi menor que o resultado do não balanceado e igual ao valor base (98,058%). Contudo os algoritmos ranqueados foram diferentes entre o não balanceado e o balanceado (Tabelas 47 e 51).

Tabela 52 – Três melhores resultados de revocação da classe positiva com todos os atributos da abordagem I no problema de novas coautorias, com balanceamento.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
meta.StackingC	F	0,5	0	0	0	0	1,942
	T	0,5	0,038	0,019	1	1	
	Avg	0,5	0,001	0	0,019	0,019	
bayes.DMNBtext	F	0,449	0	0	0	0	1,942
	T	0,449	0,038	0,019	1	1	
	Avg	0,449	0,001	0	0,019	0,019	
rules.ConjunctiveRule	F	0,626	0,549	0,993	0,38	0,128	38,91
	T	0,626	0,053	0,027	0,872	0,62	
	Avg	0,626	0,54	0,975	0,389	0,137	

Fonte: William T. Maruyama, 2015.

Quanto aos resultados referentes aos verdadeiros positivos, os melhores algoritmos classificaram todas as instâncias (ou quase todas) como positivas, registrando uma baixa precisão e baixa taxa de acerto (Tabela 52). Em relação ao não balanceado, a revocação do balanceado foi maior, mas com precisão menor (Tabela 48).

Tabela 53 – Três melhores resultados de AUC com todos os atributos da abordagem I no problema de novas coautorias, com balanceamento.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
bayes.NaiveBayes Updateable	F	0,742	0,931	0,987	0,882	0,602	87,252
	T	0,742	0,108	0,063	0,398	0,118	
	Avg	0,742	0,915	0,969	0,873	0,593	
bayes.NaiveBayes	F	0,742	0,931	0,987	0,882	0,602	87,252
	T	0,742	0,108	0,063	0,398	0,118	
	Avg	0,742	0,915	0,969	0,873	0,593	
functions.Logistic	F	0,738	0,829	0,99	0,714	0,38	71,209
	T	0,738	0,077	0,041	0,62	0,286	
	Avg	0,738	0,815	0,971	0,712	0,378	

Fonte: William T. Maruyama, 2015.

Os melhores colocados de AUC no conjunto, por sua vez, registraram valores menores comparados com os dados não balanceados (Tabelas 53 e 49). Contudo, apresentaram valores maiores na identificação da classe positiva, mas menores em precisão.

Tabela 54 – Três melhores resultados de Medida-F com todos os atributos da abordagem I no problema de novas coautorias, com balanceamento.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
meta.ThresholdSelector	F	0,738	0,922	0,988	0,864	0,54	85,636
	T	0,738	0,111	0,063	0,46	0,136	
	Avg	0,738	0,906	0,97	0,856	0,532	
bayes.NaiveBayes Updateable	F	0,742	0,931	0,987	0,882	0,602	87,252
	T	0,742	0,108	0,063	0,398	0,118	
	Avg	0,742	0,915	0,969	0,873	0,593	
bayes.NaiveBayes	F	0,742	0,931	0,987	0,882	0,602	87,252
	T	0,742	0,108	0,063	0,398	0,118	
	Avg	0,742	0,915	0,969	0,873	0,593	

Fonte: William T. Maruyama, 2015.

Com o balanceamento, os valores da Medida-F foram inferiores (Tabela 54) aos resultados obtidos sem balanceamento (Tabela 50). Esses valores podem ser explicados pelo aumento da revocação e a diminuição da precisão.

5.3.1.3 Abordagem I com atributos de domínio no problema de novas coautorias

Neste experimento foram realizados testes com o conjunto de atributos de domínio, conforme a Tabela 5, para o problema de predição de novas coautorias. A seguir, serão apresentados os resultados ranqueados de acurácia (Tabela 55), de revocação (Tabela 56), de AUC (Tabela 57) e de Medida-F (Tabela 58).

Os três valores ranqueados foram idênticos à acurácia base do problema de predição de novas coautorias (98,058%), isso porque todas (ou quase todas) as instâncias foram classificadas como negativas e os três algoritmos tiveram precisão desta classe em 0,981 (Tabela 55). Diferentemente, o 1º colocado da análise com todos os atributos sem ba-

Tabela 55 – Três melhores resultados de acurácia dos atributos de domínio da abordagem I no problema de novas coautorias, com balanceamento.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
trees.SimpleCart	F	0,5	0,99	0,981	1	1	98,058
	T	0,5	0	0	0	0	
	Avg	0,5	0,971	0,962	0,981	0,981	
trees.REPTree	F	0,575	0,99	0,981	1	0,993	98,058
	T	0,575	0,014	0,5	0,007	0	
	Avg	0,575	0,971	0,971	0,981	0,973	
trees.NBTree	F	0,586	0,99	0,981	1	1	98,058
	T	0,586	0	0	0	0	
	Avg	0,586	0,971	0,962	0,981	0,981	

Fonte: William T. Maruyama, 2015.

lanceamento (Tabela 47) apresentou acurácia um pouco superior à registrada no teste atual.

Tabela 56 – Três melhores resultados de revocação da classe positiva dos atributos de domínio da abordagem I no problema de novas coautorias, com balanceamento.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
misc.VFI	F	0,641	0,663	0,988	0,499	0,307	50,319
	T	0,641	0,051	0,027	0,693	0,501	
	Avg	0,641	0,652	0,969	0,503	0,31	
meta.Classification ViaClustering	F	0,429	0,645	0,975	0,482	0,624	47,98
	T	0,429	0,027	0,014	0,376	0,518	
	Avg	0,429	0,633	0,956	0,48	0,622	
meta.ThresholdSelector	F	0,645	0,963	0,983	0,943	0,818	92,808
	T	0,645	0,09	0,059	0,182	0,057	
	Avg	0,645	0,946	0,965	0,928	0,803	

Fonte: William T. Maruyama, 2015.

O melhor resultado de revocação da classe positiva foi registrado pelo algoritmo *VFI* (0,693), contudo este resultado apresenta-se com baixa precisão (0,027) e acurácia (50,319%). A despeito da menor revocação, a acurácia e a precisão maiores foram obtidas pelo 3º colocado no ranqueamento (Tabela 56).

O *VFI* registrou revocação menor do que o apresentado pelo teste com todos os atributos sem balanceamento, apesar dos valores próximos da precisão (Tabelas 56 e 48, respectivamente). Nesta mesma comparação, a acurácia do teste atual foi maior que a do conjunto completo.

DMNBtext foi o algoritmo que apresentou maior AUC e foi o único no ranqueamento que classificou instâncias na classe positiva (Tabela 57).

Os valores de AUC obtidos no teste com atibutos de domínio foram inferiores ao registrado com todos os atributos, apesar dos valores de acurácia e de revocação próximos entre os dois testes (Tabelas 57 e 49, respectivamente).

Tabela 57 – Três melhores resultados de AUC dos atributos de domínio da abordagem I no problema de novas coautorias, com balanceamento.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
bayes.DMNBtext	F	0,677	0,99	0,981	1	0,996	98,03
	T	0,677	0,007	0,167	0,004	0	
	Avg	0,677	0,971	0,965	0,98	0,977	
trees.ADTree	F	0,653	0,99	0,981	1	1	98,058
	T	0,653	0	0	0	0	
	Avg	0,653	0,971	0,962	0,981	0,981	
meta.LogitBoost	F	0,65	0,99	0,981	1	1	98,058
	T	0,65	0	0	0	0	
	Avg	0,65	0,971	0,962	0,981	0,981	

Fonte: William T. Maruyama, 2015.

Tabela 58 – Três melhores resultados da Medida-F dos atributos de domínio da abordagem I no problema de novas coautorias, com balanceamento.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
bayes.NaiveBayes Updateable	F	0,641	0,982	0,982	0,982	0,898	96,45
	T	0,641	0,101	0,099	0,102	0,018	
	Avg	0,641	0,965	0,965	0,964	0,881	
bayes.NaiveBayes	F	0,641	0,982	0,982	0,982	0,898	96,45
	T	0,641	0,101	0,099	0,102	0,018	
	Avg	0,641	0,965	0,965	0,964	0,881	
meta.ThresholdSelector	F	0,645	0,963	0,983	0,943	0,818	92,808
	T	0,645	0,09	0,059	0,182	0,057	
	Avg	0,645	0,946	0,965	0,928	0,803	

Fonte: William T. Maruyama, 2015.

Os melhores valores para a Medida-F foram obtidos pelos algoritmos *Naive Bayes Updateable* e *Naive Bayes*, ambos com 0,101. Os valores das outras métricas obtidos por estes algoritmos também são semelhantes neste teste (Tabela 58).

Comparando, o conjunto total de atributos apresentou resultados melhores (Tabela 50) em relação ao teste com subconjunto de atributos de domínio apresentado na Tabela 58.

5.3.1.4 Abordagem I com atributos estruturais no problema de novas coautorias

Este teste foi realizado a partir dos atributos estruturais da rede, conforme descrito na Tabela 5. Do mesmo modo que os testes anteriores, como resultado são apresentados os três melhores resultados das métricas acurácia (Tabela 59), revocação (Tabela 60), AUC (Tabela 61) e Medida-F (Tabela 62).

Os três melhores resultados em acurácia são iguais aos valores base do problema (98,058%). Isso devido à classificação de todas as instâncias como sendo negativas (Tabela 59). Este resultado é um pouco inferior ao apresentado pelo experimento com todos

Tabela 59 – Três melhores resultados de acurácia dos atributos estruturais da abordagem I no problema de novas coautorias.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
trees.SimpleCart	F	0,5	0,99	0,981	1	1	98,058
	T	0,5	0	0	0	0	
	Avg	0,5	0,971	0,962	0,981	0,981	
trees.NBTree	F	0,629	0,99	0,981	1	1	98,058
	T	0,629	0	0	0	0	
	Avg	0,629	0,971	0,962	0,981	0,981	
trees.FT	F	0,671	0,99	0,981	1	1	98,058
	T	0,671	0	0	0	0	
	Avg	0,671	0,971	0,962	0,981	0,981	

Fonte: William T. Maruyama, 2015.

os atributos sem balanceamento e com os atributos de domínio, inclusive quanto à precisão e revocação da classe positiva (Tabelas 47 e 55, respectivamente).

Tabela 60 – Três melhores resultados de revocação da classe positiva dos atributos estruturais da abordagem I no problema de novas coautorias.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
misc.VFI	F	0,641	0,782	0,988	0,647	0,387	64,647
	T	0,641	0,063	0,033	0,613	0,353	
	Avg	0,641	0,768	0,97	0,646	0,386	
meta.Classification ViaClustering	F	0,643	0,897	0,987	0,822	0,536	81,526
	T	0,643	0,089	0,049	0,464	0,178	
	Avg	0,643	0,882	0,969	0,815	0,53	
bayes.BayesNet	F	0,696	0,908	0,987	0,84	0,573	83,213
	T	0,696	0,09	0,05	0,427	0,16	
	Avg	0,696	0,892	0,968	0,832	0,565	

Fonte: William T. Maruyama, 2015.

Quanto aos resultados da revocação da classe positiva, o maior valor foi registrado por *VFI* (0,613), contudo, dentre os três ranqueados, ele foi o que de menor taxa de acerto (64,647%, Tabela 60).

Comparativamente, o presente teste também apresentou desempenho inferior aos registrados nos conjuntos completo e de atributos de domínio, contudo pode-se destacar que a acurácia e a precisão do 1º colocado é maior do que o registrado na mesma posição nesses dois experimentos (Tabelas 60, 48 e 56, respectivamente).

Tabela 61 – Três melhores resultados de AUC dos atributos estruturais da abordagem I no problema de novas coautorias.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
meta.LogitBoost	F	0,744	0,99	0,981	1	0,996	98,03
	T	0,744	0,007	0,167	0,004	0	
	Avg	0,744	0,971	0,965	0,98	0,977	
meta.Bagging	F	0,74	0,99	0,981	1	1	98,058
	T	0,74	0	0	0	0	
	Avg	0,74	0,971	0,962	0,981	0,981	
trees.LADTree	F	0,703	0,99	0,981	0,998	0,989	97,931
	T	0,703	0,02	0,125	0,011	0,002	
	Avg	0,703	0,971	0,964	0,979	0,97	

Fonte: William T. Maruyama, 2015.

O algoritmo *Logit Boost* foi o que apresentou maior AUC (0,744), valor este menor que o registrado com todos os atributos sem balanceamento, mas maior que o obtido pelos atributos de domínio (Tabelas 61, 49 e 57, respectivamente).

Tabela 62 – Três melhores resultados da Medida-F dos atributos estruturais da abordagem I no problema de novas coautorias.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
bayes.NaiveBayesSimple	F	0,65	0,955	0,985	0,927	0,734	91,376
	T	0,65	0,107	0,067	0,266	0,073	
	Avg	0,65	0,938	0,967	0,914	0,721	
bayes.NaiveBayes Updateable	F	0,647	0,949	0,985	0,916	0,715	90,405
	T	0,647	0,103	0,063	0,285	0,084	
	Avg	0,647	0,933	0,967	0,904	0,703	
bayes.NaiveBayes	F	0,647	0,949	0,985	0,916	0,715	90,405
	T	0,647	0,103	0,063	0,285	0,084	
	Avg	0,647	0,933	0,967	0,904	0,703	

Fonte: William T. Maruyama, 2015.

Com 0,107, o algoritmo *Naive Bayes Simple* foi o que registrou melhor valor em Medida-F (Tabela 62). Valor este menor do que os apresentados pelos testes com o conjunto total de atributos sem balanceamento e com os atributos de domínio (Tabelas 50 e 58, respectivamente).

5.3.1.5 Abordagem I com seleção de atributo no problema de novas coautorias

Da mesma forma que no problema geral de predição, foram realizados testes com subconjuntos formados por algoritmos de seleção de atributos mais relevantes para a predição. A Tabela 63 contém o resultado da execução de todos os seletores de atributos disponíveis no Weka que retornaram subconjuntos não vazios. Os algoritmos de seleção utilizados apenas indicam quais atributos foram selecionados (marcados por "x").

Para analisar os resultados dos subconjuntos formados, são apresentados os respectivos ranqueamentos das métricas acurácia (Tabela 64), revocação da classe T (Tabela 65), AUC (Tabela 66) e Medida-F (Tabela 67).

Foram formados 10 subconjuntos, sendo o Subconjunto 4 o menor (1 atributo) e o Subconjunto 6 o maior (18 atributos). Dois subconjuntos com algoritmos de seleção diferentes retornaram os mesmos atributos (Subconjuntos 2 e 9). Além disso, pode-se observar que o atributo "Conferências passado" está presente em todos os subconjuntos, contudo 9 atributos ("Periódicos presente", "Conferências presente", "Orientação passado", "Artigos em periódicos 1", "Artigos em periódicos 2", "Artigos em anais 2", "HDI", "PA" e "Katz 0,05") não foram selecionados para nenhum subconjunto formado (Tabela 63).

Tabela 63 – Subconjuntos obtidos com seleção de características da abordagem I no problema de novas coautorias.

	Subconjuntos									
	1	2	3	4	5	6	7	8	9	10
Periódicos passado	x				x	x	x	x		
Conferências passado	x	x	x	x	x	x	x	x	x	x
Periódicos presente										
Conferências presente										
Orientação passado										
Orientação presente	x	x	x		x	x	x	x	x	x
Orientação em andamento	x	x	x		x	x	x	x	x	x
Orientadores em comum						x				
Orientandos em comum						x				
Passado e presente CN	x	x	x		x	x	x	x	x	x
Programas em comum	x	x	x		x	x	x	x	x	x
Artigos em periódicos 1										
Artigos em anais 1						x				
Artigos em periódicos 2										
Artigos em anais 2										
Presente CN						x				
SA						x				
JC					x					
AA					x	x	x			
RA	x	x	x		x	x	x	x	x	x
SO			x				x			x
HPI							x			
HDI										
LHN					x	x	x			
PA										
KATZ 0,05										
KATZ 0,005	x	x	x		x		x	x	x	x
KATZ 0,0005						x				
Subáreas em comum					x	x	x			
Distância geográfica	x	x	x		x	x	x	x	x	x
Distância no grafo (SP)			x		x	x	x			x

Nota: Os números dos subcojuntos correspondem aos algoritmos testados. O primeiro algoritmo é o de seleção seguido do(s) método(s) de busca utilizado(s):
(1) *CfsSubsetEval* / *GreedyStepwise*
(2) *CfsSubsetEval* / *BestFirst* e *LinearForwardSelection*
(3) *CfsSubsetEval* / *GeneticSearch*
(4) *ConsistencySubsetEval* / *GreedyStepwise*
(5) *ConsistencySubsetEval* / *BestFirst*
(6) *ConsistencySubsetEval* / *GeneticSearch*
(7) *ConsistencySubsetEval* / *LinearForwardSelection*
(8) *FilteredSubsetEval* / *GreedyStepwise*
(9) *FilteredSubsetEval* / *BestFirst* e *LinearForwardSelection*
(10) *FilteredSubsetEval* / *GeneticSearch*

Fonte: William T. Maruyama, 2015.

Tabela 64 – Três melhores resultados de acurácia com seleção de atributos da abordagem I no problema de novas coautorias.

Subconjunto	Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
3	meta.Bagging	F	0,708	0,99	0,981	1	0,993	98,065
		T	0,708	0,014	0,667	0,007	0	
		Avg	0,708	0,971	0,975	0,981	0,973	
10	meta.Bagging	F	0,709	0,99	0,981	1	0,993	98,065
		T	0,709	0,014	0,667	0,007	0	
		Avg	0,709	0,971	0,975	0,981	0,973	
2	functions.Multilayer Perceptron	F	0,726	0,99	0,981	1	0,996	98,058
		T	0,726	0,007	0,5	0,004	0	
		Avg	0,726	0,971	0,971	0,981	0,977	

Fonte: William T. Maruyama, 2015.

Os valores de acurácia foram, conforme o ranqueamento, maiores ou iguais ao valor base (98,058%) do problema de novas coautorias, sendo os Subconjuntos 3 e 10 com os maiores valores (98,065%, Tabela 64). Semelhantemente ao registrado na análise com todos os atributos sem balanceamento (Tabela 47), todas (ou quase todas) as instâncias foram classificadas como negativas.

Tabela 65 – Três melhores resultados de revocação da classe positiva com seleção de atributos da abordagem I no problema de novas coautorias.

Subconjunto	Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
3	misc.VFI	F	0,713	0,582	0,993	0,411	0,139	42,014
		T	0,713	0,055	0,028	0,861	0,589	
		Avg	0,713	0,572	0,975	0,42	0,147	
2 e 9	misc.VFI	F	0,691	0,581	0,993	0,41	0,139	41,886
		T	0,691	0,054	0,028	0,861	0,59	
		Avg	0,691	0,57	0,975	0,419	0,147	
5	misc.VFI	F	0,721	0,58	0,993	0,409	0,146	41,794
		T	0,721	0,054	0,028	0,854	0,591	
		Avg	0,721	0,569	0,974	0,418	0,155	

Fonte: William T. Maruyama, 2015.

A revocação e a precisão dos subconjuntos foi igual nos três melhores ranqueados (0,861 e 0,028, respectivamente), variando somente a acurácia de cada subconjunto (Tabela 65). Deste modo, o Subconjunto 3 foi o que apresentou melhor resultado. Quanto à comparação com o teste do conjunto completo sem balanceamento do treinamento (Tabela 48), a seleção de atributos propiciou uma leve melhora no desempenho de revocação, contudo a precisão continuou baixa

Tabela 66 – Três melhores resultados de AUC com seleção de atributos da abordagem I no problema de novas coautorias.

Subconjunto	Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
6	meta.Classification	F	0,758	0,99	0,981	1	0,989	98,044
	ViaRegression	T	0,758	0,021	0,375	0,011	0	
		Avg	0,758	0,971	0,969	0,98	0,97	
5	bayes.NaiveBayes	F	0,745	0,963	0,985	0,943	0,745	92,928
	Updateable	T	0,745	0,123	0,081	0,255	0,057	
		Avg	0,745	0,947	0,967	0,929	0,731	
7	bayes.NaiveBayes	F	0,744	0,959	0,985	0,935	0,719	92,205
	Updateable	T	0,743	0,123	0,079	0,281	0,065	
		Avg	0,744	0,943	0,967	0,922	0,706	

Fonte: William T. Maruyama, 2015.

O subconjunto que apresentou melhor valor de AUC - Subconjunto 6 - também registrou o menor valor de revocação da classe positiva e a maior precisão e, consequentemente acurácia, dentre os três ranqueados (Tabela 66). Este primeiro colocado apresentou AUC um pouco maior do que o registrado na análise do conjunto completo de atributos sem balanceamento (Tabela 49), assim como a sua acurácia foi maior.

Tabela 67 – Três melhores resultados da Medida-F com seleção de atributos da abordagem I no problema de novas coautorias.

Subconjunto	Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
1	bayes.NaiveBayes Updateable	F	0,738	0,975	0,984	0,966	0,788	95,174
		T	0,738	0,146	0,111	0,212	0,034	
		Avg	0,738	0,959	0,967	0,952	0,774	
2	bayes.NaiveBayes Updateable	F	0,734	0,976	0,984	0,968	0,796	95,366
		T	0,734	0,146	0,114	0,204	0,032	
		Avg	0,734	0,96	0,967	0,954	0,781	
9	bayes.NaiveBayes Updateable	F	0,734	0,976	0,984	0,968	0,796	95,366
		T	0,734	0,146	0,114	0,204	0,032	
		Avg	0,734	0,96	0,967	0,954	0,781	

Fonte: William T. Maruyama, 2015.

Os resultados alcançados com os subconjuntos da Tabela 67, quanto ao desempenho da Medida-F, foram maiores que os obtidos com o conjunto total de atributos (Tabela 50). Em comparação ao conjunto total de atributos, ocorreu uma aproximação entre os valores de precisão e revocação, já que a precisão aumentou e a revocação diminuiu.

5.3.1.6 Abordagem I com atributos individuais no problema de novas coautorias

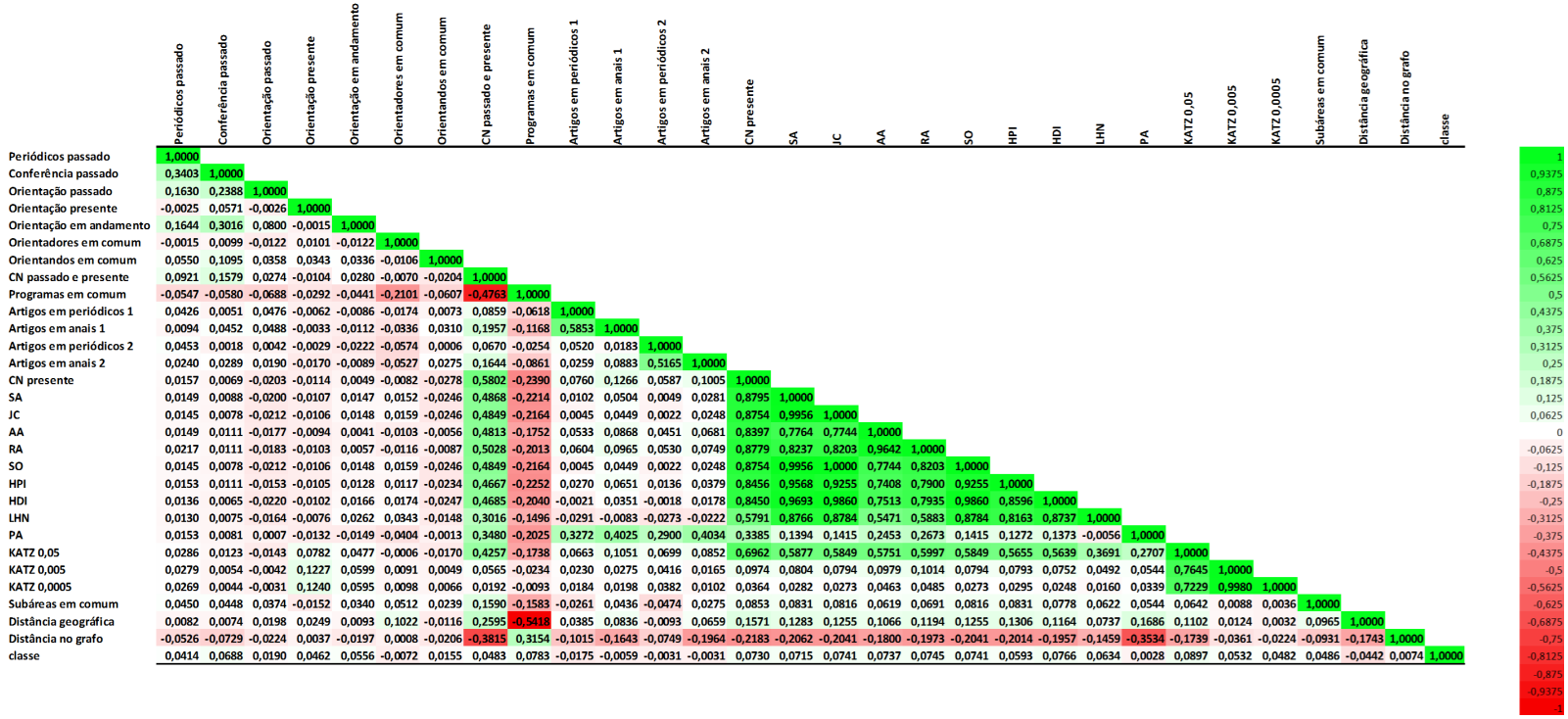
Assim como no experimento da subseção 5.2.1.6, foi realizado o mesmo procedimento de análise para o problema de novas coautorias. Logo, a Tabela 68 apresenta a ordenação dos atributos e a Figura 12 contém a matriz de correlação entre a classe e os demais atributos.

É possível observar que os atributos melhor ranqueados, conforme valores da mediana, são “Katz 0,005” (2), “Programas em comum” (3) e “Conferências passado” (5). De maneira oposta, os piores são “Artigos em periódicos 2” (28), “PA” (27) e “Artigos em anais 1” (27, Tabela 68).

Os valores de correlação entre os atributos e a classe (última linha) não apresentaram valores maiores que 0,1 ou menores que -0,1, demonstrando que a relação entre eles não é muito relevante, mas a maior parte apresenta uma correlação positiva. Pode-se destacar que, novamente, o atributo “Programas em comum” é o que apresenta maior correlação negativa com os outros atributos, principalmente com a “Distância geográfica” (Figura 12).

A seguir, são apresentados os resultados da classificação ao utilizar cada um dos atributos individualmente. Os resultados de cada atributo foram ranqueados para cada métrica analisada, contudo serão apresentados apenas os três melhores atributos em relação a cada métrica.

Figura 12 – Matriz de correlação dos atributos individuais no problema de novas coautorias.



	Periódicos passado	Conferência passado	Orientação passado	Orientação presente	Orientação em andamento	Orientadores em comum	Orientandos em comum	CN passado e presente	Programas em comum	Artigos em periódicos 1	Artigos em anais 1	Artigos em periódicos 2	Artigos em anais 2	CN presente	SA	JC	AA	RA	SO	HPI	HDI	LHN	PA	KATZ 0,05	KATZ 0,005	KATZ 0,0005	Subáreas em comum	Distância geográfica	Distância no grafo	classe
Periódicos passado	1,0000																													
Conferência passado	0,3403	1,0000																												
Orientação passado	0,1630	0,2388	1,0000																											
Orientação presente	-0,0025	0,0571	-0,0026	1,0000																										
Orientação em andamento	0,1644	0,3016	0,0800	-0,0015	1,0000																									
Orientadores em comum	-0,0015	0,0099	-0,0122	0,0101	-0,0122	1,0000																								
Orientandos em comum	0,0550	0,1095	0,0358	0,0343	0,0336	-0,0106	1,0000																							
CN passado e presente	0,0921	0,1579	0,0274	-0,0104	0,0280	-0,0070	-0,0204	1,0000																						
Programas em comum	-0,0547	-0,0580	-0,0688	-0,0292	-0,0441	-0,2101	-0,0607	-0,4763	1,0000																					
Artigos em periódicos 1	0,0426	0,0051	0,0476	-0,0062	-0,0086	-0,0174	0,0073	0,0859	-0,0618	1,0000																				
Artigos em anais 1	0,0094	0,0452	0,0488	-0,0033	-0,0112	-0,0336	0,0310	0,1957	-0,1168	0,5853	1,0000																			
Artigos em periódicos 2	0,0453	0,0018	0,0042	-0,0029	-0,0222	-0,0574	0,0006	0,0670	-0,0254	0,0520	0,0183	1,0000																		
Artigos em anais 2	0,0240	0,0289	0,0190	-0,0170	-0,0089	-0,0527	0,0275	0,1644	-0,0861	0,0259	0,0883	0,5165	1,0000																	
CN presente	0,0157	0,0069	-0,0203	-0,0114	0,0049	-0,0082	-0,0278	0,5802	-0,2390	0,0760	0,1266	0,0587	0,1005	1,0000																
SA	0,0149	0,0088	-0,0200	-0,0107	0,0147	0,0152	-0,0246	0,4868	-0,2214	0,0102	0,0504	0,0049	0,0281	0,8795	1,0000															
JC	0,0145	0,0078	-0,0212	-0,0106	0,0148	0,0159	-0,0246	0,4849	-0,2164	0,0045	0,0449	0,0022	0,0248	0,8754	0,9956	1,0000														
AA	0,0149	0,0111	-0,0177	-0,0094	0,0041	-0,0103	-0,0056	0,4813	-0,1752	0,0533	0,0868	0,0451	0,0681	0,8397	0,7764	0,7744	1,0000													
RA	0,0217	0,0111	-0,0183	-0,0103	0,0057	-0,0116	-0,0087	0,5028	-0,2013	0,0604	0,0965	0,0530	0,0749	0,8779	0,8237	0,8203	0,9642	1,0000												
SO	0,0145	0,0078	-0,0212	-0,0106	0,0148	0,0159	-0,0246	0,4849	-0,2164	0,0045	0,0449	0,0022	0,0248	0,8754	0,9956	1,0000	0,7744	0,8203	1,0000											
HPI	0,0153	0,0111	-0,0153	-0,0105	0,0128	0,0117	-0,0234	0,4667	-0,2252	0,0270	0,0651	0,0136	0,0379	0,8456	0,9568	0,9255	0,7408	0,7900	0,9255	1,0000										
HDI	0,0136	0,0065	-0,0220	-0,0102	0,0166	0,0174	-0,0247	0,4685	-0,2040	-0,0021	0,0351	-0,0018	0,0178	0,8450	0,9693	0,9860	0,7513	0,7935	0,9860	0,8596	1,0000									
LHN	0,0130	0,0075	-0,0164	-0,0076	0,0262	0,0343	-0,0148	0,3016	-0,1496	-0,0291	-0,0083	-0,0273	-0,0222	0,5791	0,8766	0,8784	0,5471	0,5883	0,8784	0,8163	0,8737	1,0000								
PA	0,0153	0,0081	0,0007	-0,0132	-0,0149	-0,0404	-0,0013	0,3480	-0,2025	0,3272	0,4025	0,2900	0,4034	0,3385	0,1394	0,1415	0,2453	0,2673	0,1415	0,1272	0,1373	-0,0056	1,0000							
KATZ 0,05	0,0286	0,0123	-0,0143	0,0782	0,0477	-0,0006	-0,0170	0,4257	-0,1738	0,0663	0,1051	0,0699	0,0852	0,6962	0,5877	0,5849	0,5751	0,5997	0,5849	0,5655	0,5639	0,3691	0,2707	1,0000						
KATZ 0,005	0,0279	0,0054	-0,0042	0,1227	0,0599	0,0091	0,0049	0,0565	-0,0234	0,0230	0,0275	0,0416	0,0165	0,0974	0,0804	0,0794	0,0979	0,1014	0,0794	0,0793	0,0752	0,0492	0,0544	0,7645	1,0000					
KATZ 0,0005	0,0269	0,0044	-0,0031	0,1240	0,0595	0,0098	0,0066	0,0192	-0,0093	0,0184	0,0198	0,0382	0,0102	0,0364	0,0282	0,0273	0,0463	0,0485	0,0273	0,0295	0,0248	0,0160	0,0339	0,7229	0,9980	1,0000				
Subáreas em comum	0,0450	0,0448	0,0374	-0,0152	0,0340	0,0512	0,0239	0,1590	-0,1583	-0,0261	0,0436	-0,0474	0,0275	0,0853	0,0831	0,0816	0,0619	0,0691	0,0816	0,0831	0,0778	0,0622	0,0544	0,0642	0,0088	0,0036	1,0000			
Distância geográfica	0,0082	0,0074	0,0198	0,0249	0,0093	0,1022	-0,0116	0,2595	-0,5418	0,0385	0,0836	-0,0093	0,0659	0,1571	0,1283	0,1255	0,1066	0,1194	0,1255	0,1306	0,1164	0,0737	0,1686	0,1102	0,0124	0,0032	0,0965	1,0000		
Distância no grafo	-0,0526	-0,0729	-0,0224	0,0037	-0,0197	0,0008	-0,0206	-0,3815	0,3154	-0,1015	-0,1643	-0,0749	-0,1964	-0,2183	-0,2062	-0,2041	-0,1800	-0,1973	-0,2041	-0,2014	-0,1957	-0,1459	-0,3534	-0,1739	-0,0361	-0,0224	-0,0931	-0,1743	1,0000	
classe	0,0414	0,0688	0,0190	0,0462	0,0556	-0,0072	0,0155	0,0483	0,0783	-0,0175	-0,0059	-0,0031	-0,0031	0,0730	0,0715	0,0741	0,0737	0,0745	0,0741	0,0593	0,0766	0,0634	0,0028	0,0897	0,0532	0,0482	0,0486	-0,0442	0,0074	1,0000

Fonte: William T. Maruyama, 2015.

Tabela 68 – Ranqueamento dos atributos do problema de novas coautorias.

	ChiSquaredAttributeEval	*FilteredAttributeEval*	*GainRatioAttributeEval*	*InfoGainAttributeEval*	*OneRAttributeEval*	*ReliefFAttributeEval*	*SymmetricalUncertAttributeEval*	Mediana
KATZ 0.005	1	2	7	2	31	27	2	2
Programas em comum	4	1	9	1	4	1	3	3
Conferências passado	5	16	5	16	1	22	5	5
AA	2	4	10	4	26	20	6	6
RA	3	7	3	7	25	18	1	7
Distância geográfica	7	3	17	3	10	3	11	7
KATZ 0.05	8	5	11	5	30	23	7	8
KATZ 0.0005	9	6	12	6	20	28	8	9
HDI	10	9	13	9	23	14	12	12
Distância no grafo (SP)	18	17	6	17	3	12	9	12
CN presente	6	8	16	8	16	24	13	13
CN passado e presente	14	15	8	15	29	9	10	14
JC	12	10	14	10	18	17	14	14
SA	15	14	19	14	17	15	17	15
SO	11	11	15	11	28	16	15	15
HPI	16	13	20	13	24	11	18	16
LHN	13	12	18	12	22	21	16	16
Orientação em andamento	17	19	2	19	9	25	4	17
Orientação presente	19	21	1	21	11	31	19	19
Periódicos passado	21	20	4	20	27	26	20	20
Subáreas em comum	20	18	21	18	21	7	21	20
Periódicos presente	23	22	22	22	12	29	22	22
Artigos em peridicos 1	26	23	30	23	2	6	23	23
Orientação passado	24	24	23	24	14	19	24	24
Conferências presente	22	25	24	25	13	30	25	25
Artigos em anais 2	29	26	28	26	15	2	26	26
Orientadores em comum	27	29	26	29	6	8	29	27
Artigos em anais 1	25	30	31	30	5	4	27	27
PA	31	27	25	27	19	10	30	27
Orientandos em comum	30	28	27	28	8	13	31	28
Artigos em periódicos 2	28	31	29	31	7	5	28	28

Fonte: William T. Maruyama, 2015.

Tabela 69 – Três melhores atributos em relação a acurácia da abordagem I no problema de novas coautorias.

Atributo	Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
KATZ 0,05	functions.Logistic	F	0,654	0,99	0,981	1	0,996	98,065
		T	0,654	0,007	1	0,004	0	
		Avg	0,654	0,971	0,981	0,981	0,977	
CN Presente	functions.Logistic	F	0,652	0,99	0,981	1	0,993	98,065
		T	0,652	0,014	0,667	0,007	0	
		Avg	0,652	0,971	0,975	0,981	0,973	
KATZ 0,0005	bayes.NaiveBayesSimple	F	0,646	0,99	0,981	1	0,996	98,065
		T	0,646	0,007	1	0,004	0	
		Avg	0,646	0,971	0,981	0,981	0,977	

Fonte: William T. Maruyama, 2015.

Os resultados obtidos de acurácia neste teste foram um pouco melhores do que os obtidos com o conjunto completo e com os subconjuntos de atributos testados anteriormente

(Tabelas 69, 47 e 64, respectivamente). Isso porque todos os três ranqueados apresentaram acurácia maior que o valor base do problema (98,058%).

Tabela 70 – Três melhores atributos em relação a revocação da classe positiva da abordagem I no problema de novas coautorias.

Atributo	Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
PA	meta.ThresholdSelector	F	0,567	0	0	0	0	1,942
		T	0,567	0,038	0,019	1	1	
		Avg	0,567	0,001	0	0,019	0,019	
Artigos em conferência 2	misc.VFI	F	0,5	0	0	0	0	1,942
		T	0,5	0,038	0,019	1	1	
		Avg	0,5	0,001	0	0,019	0,019	
Periódicos presente	misc.VFI	F	0,5	0	0	0	0	1,942
		T	0,5	0,038	0,019	1	1	
		Avg	0,5	0,001	0	0,019	0,019	

Fonte: William T. Maruyama, 2015.

A revocação dos atributos individuais (Tabela 70) foi superior que a do conjunto completo (Tabela 48) e dos subconjuntos selecionados (Tabela 65), isso devido a todas as instâncias serem classificadas como da classe positiva. Entretanto, a precisão atual foi menor do que a dos mesmos testes citados.

Tabela 71 – Três melhores atributos em relação a AUC da abordagem I no problema de novas coautorias.

Atributo	Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
KATZ 0,05	lazy.LWL	F	0,673	0,99	0,981	1	1	98,058
		T	0,673	0	0	0	0	
		Avg	0,673	0,971	0,962	0,981	0,981	
KATZ 0,005	meta.Classification ViaRegression	F	0,663	0,99	0,981	1	1	98,058
		T	0,663	0	0	0	0	
		Avg	0,663	0,971	0,962	0,981	0,981	
AA	functions.RBFNetwork	F	0,656	0,99	0,981	1	1	98,058
		T	0,656	0	0	0	0	
		Avg	0,656	0,971	0,962	0,981	0,981	

Fonte: William T. Maruyama, 2015.

Os melhores desempenhos em AUC foram de três atributos estruturais, sendo que foram, comparativamente, inferiores aos registrado no experimento com o conjunto completo e com os subconjuntos (Tabela 71, 49 e 66, respectivamente).

Tabela 72 – Três melhores atributos em relação a Medida-F da abordagem I no problema de novas coautorias.

Atributo	Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
CN presente	bayes.NaiveBayesSimple	F	0,617	0,973	0,984	0,962	0,796	94,763
		T	0,617	0,132	0,097	0,204	0,038	
		Avg	0,617	0,957	0,967	0,948	0,781	
KATZ 0,05	meta.Classification ViaClustering	F	0,557	0,978	0,983	0,972	0,858	95,635
		T	0,557	0,112	0,093	0,142	0,028	
		Avg	0,557	0,961	0,966	0,956	0,842	
KATZ 0,005	meta.ThresholdSelector	F	0,663	0,979	0,983	0,974	0,865	95,812
		T	0,663	0,111	0,095	0,135	0,026	
		Avg	0,663	0,962	0,965	0,958	0,849	

Fonte: William T. Maruyama, 2015.

O valor da Medida-F do 1º ranqueado no teste atual (Tabela 72) foi superior ao obtido com o conjunto total de atributos não balanceados (Tabela 50). Contudo, os resultados da seleção de atributos foram superiores (Tabela 67).

5.3.1.7 Abordagem I com atributos individuais e balanceado no problema de novas coautorias

Nesse experimento foi realizado o balanceamento do conjunto de treinamento e cada atributo (Tabela 68) foi utilizado individualmente para classificação. Assim como no experimento anterior, são apresentados os resultados da classificação ao utilizar cada um dos atributos individualmente, ranqueados para cada métrica analisada. Isto é, serão apresentados os melhores atributos em relação a acurácia (Tabela 73), a revocação da classe positiva (Tabela 74), a AUC (Tabela 75) e a Medida-F (Tabela 76).

Tabela 73 – Três melhores atributos em relação à acurácia da abordagem I no problema de novas coautorias, com balanceamento.

Atributo	Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
KATZ 0,0005	lazy.KStar	F	0,646	0,99	0,981	1	0,996	98,065
		T	0,646	0,007	1	0,004	0	
		Avg	0,646	0,971	0,981	0,981	0,977	
KATZ 0,005	bayes.NaiveBayesSimple	F	0,64	0,99	0,981	1	0,996	98,065
		T	0,64	0,007	1	0,004	0	
		Avg	0,64	0,971	0,981	0,981	0,977	
KATZ 0,05	functions.LibSVM	F	0,502	0,99	0,981	1	0,996	98,065
		T	0,502	0,007	1	0,004	0	
		Avg	0,502	0,971	0,981	0,981	0,977	

Fonte: William T. Maruyama, 2015.

As três variações de Katz foram as que apresentaram melhores desempenhos em acurácia no presente teste, sendo todos maiores que o valor base do problema (98,058%, Tabela 73). Além disso, com exceção do 2º atributo, o restante dos atributos foram semelhantes aos registrados no teste com os atributos individuais não balanceado (Tabela 69). Nesta mesma linha comparativa, os valores de acurácia entre os dois foram iguais também, diferenciando um pouco somente quanto à revocação e precisão.

Já os resultados de revocação do conjunto balanceado apresentam que as instâncias foram todas classificadas como classe positiva (Tabela 74), assim como no teste não balanceado (Tabela 70). Além disso, os mesmos valores de precisão e acurácia foram registrados entre os dois experimentos.

Os valores de AUC apresentados na Tabela 75 foram iguais entre este conjunto balanceado e àqueles obtidos utilizando-se como treinamento o conjunto não balanceado

Tabela 74 – Três melhores atributos em relação à revocação da classe positiva da abordagem I no problema de novas coautorias, com balanceamento.

Atributo	Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
CN	bayes.DMNBtext	F	0,646	0	0	0	0	1,942
		T	0,646	0,038	0,019	1	1	
		Avg	0,646	0,001	0	0,019	0,019	
SA	bayes.DMNBtext	F	0,646	0	0	0	0	1,942
		T	0,646	0,038	0,019	1	1	
		Avg	0,646	0,001	0	0,019	0,019	
JC	bayes.DMNBtext	F	0,646	0	0	0	0	1,942
		T	0,646	0,038	0,019	1	1	
		Avg	0,646	0,001	0	0,019	0,019	

Fonte: William T. Maruyama, 2015.

Tabela 75 – Três melhores atributos em relação à AUC da abordagem I no problema de novas coautorias, com balanceamento.

Atributo	Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
KATZ 0,05	lazy.KStar	F	0,674	0,838	0,987	0,728	0,474	72,378
		T	0,674	0,069	0,037	0,526	0,272	
		Avg	0,674	0,823	0,969	0,724	0,471	
KATZ 0,005	meta.ThresholdSelector	F	0,663	0,888	0,987	0,806	0,518	79,974
		T	0,663	0,085	0,047	0,482	0,194	
		Avg	0,663	0,872	0,969	0,8	0,512	
AA	meta.ThresholdSelector	F	0,656	0,983	0,982	0,983	0,916	96,592
		T	0,656	0,087	0,091	0,084	0,017	
		Avg	0,656	0,965	0,965	0,966	0,899	

Fonte: William T. Maruyama, 2015.

(Tabela 71). Entretanto, o desempenho mensurado pela revocação da classe positiva e a precisão, bem como a acurácia, são maiores nos atributos individuais utilizando o conjunto balanceado.

Tabela 76 – Três melhores atributos em relação à Medida-F da abordagem I no problema de novas coautorias, com balanceamento.

Atributo	Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
KATZ 0,05	meta.Dagging	F	0,594	0,979	0,983	0,975	0,865	95,897
		T	0,594	0,113	0,098	0,135	0,025	
		Avg	0,594	0,962	0,966	0,959	0,849	
Distância geográfica	trees.DecisionStump	F	0,582	0,966	0,984	0,948	0,785	93,389
		T	0,582	0,112	0,076	0,215	0,052	
		Avg	0,582	0,949	0,966	0,934	0,77	
KATZ 0,005	bayes.BayesianLogistic Regression	F	0,558	0,976	0,983	0,969	0,854	95,323
		T	0,558	0,108	0,086	0,146	0,031	
		Avg	0,558	0,959	0,965	0,953	0,838	

Fonte: William T. Maruyama, 2015.

Os valores referentes à Medida-F do teste atual (Tabela 76) foram próximos, mas menores, em relação ao conjunto total de atributos balanceados (Tabela 54). Quanto ao conjunto de atributos individuais não balanceado (Tabela 72), a Medida-F registrada após o balanceamento apresenta desempenho inferior.

5.3.2 Abordagem II

As quantidades de instâncias por classe do conjunto utilizada na abordagem II não balanceada do problema de novas coautorias pode ser observada na Tabela 77. Para o conjunto de treinamento balanceado, o valor da classe minoritária (classe T) é igualado à quantidade da classe majoritária (classe F). Pode-se observar que, caso todas as instâncias fossem classificadas como negativas (valor base), a acurácia seria de 98,058%.

Tabela 77 – Quantidade de instâncias da abordagem II no problema de novas coautorias.

	Classe	
	F	T
Conjunto de treinamento	10537	597
Conjunto de teste	13838	274

Fonte: William T. Maruyama, 2015.

5.3.2.1 Abordagem II com todos atributos no problema de novas coautorias

São apresentados a seguir os algoritmos que obtiveram as maiores taxas conforme acurácia (Tabela 78), revocação (Tabela 79), AUC (Tabela 80) e Medida-F (Tabela 81), ao se analisar todos os atributos sem o balanceamento do conjunto de treinamento.

Tabela 78 – Três melhores resultados de acurácia com todos os atributos da abordagem II no problema de novas coautorias.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
meta.RotationForest	F	0,692	0,99	0,981	1	0,996	98,065
	T	0,692	0,007	1	0,004	0	
	Avg	0,692	0,971	0,981	0,981	0,977	
trees.ADTree	F	0,684	0,99	0,981	1	1	98,058
	T	0,684	0	0	0	0	
	Avg	0,684	0,971	0,962	0,981	0,981	
meta.RandomSubSpace	F	0,68	0,99	0,981	1	1	98,058
	T	0,68	0	0	0	0	
	Avg	0,68	0,971	0,962	0,981	0,981	

Fonte: William T. Maruyama, 2015.

Nessa abordagem, o algoritmo que apresentou maior acurácia foi o *Rotation Forest* (98,065%), sendo ele superior ao valor base (98,058%). A diferença observada é que ele foi o único a classificar algumas instâncias na classe T (0,004) e obteve alta precisão (igual a 1) nessa classificação. As outras acurácias foram iguais ao valor base, pois todas as instâncias foram classificadas como da classe negativa (Tabela 78).

Comparando os resultados entre as abordagens para o mesmo problema, este é um resultado parecido, o qual difere somente que dois algoritmos classificaram classes positivas na abordagem I (Tabela 47).

Tabela 79 – Três melhores resultados de revocação da classe positiva com todos os atributos da abordagem II no problema de novas coautorias.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
misc.VFI	F	0,669	0,779	0,988	0,643	0,38	64,243
	T	0,669	0,063	0,033	0,62	0,357	
	Avg	0,669	0,765	0,97	0,642	0,379	
bayes.NaiveBayes Updateable	F	0,677	0,949	0,986	0,914	0,679	90,271
	T	0,677	0,114	0,069	0,321	0,086	
	Avg	0,677	0,932	0,968	0,903	0,667	
bayes.NaiveBayes	F	0,677	0,949	0,986	0,914	0,679	90,271
	T	0,677	0,114	0,069	0,321	0,086	
	Avg	0,677	0,932	0,968	0,903	0,667	

Fonte: William T. Maruyama, 2015.

VFI foi o algoritmo que apresentou maior revocação (0,62), contudo obteve menor precisão e acurácia dentre os três ranqueados (Tabela 79). Este teste registrou menores valores de revocação da classe T do que os apresentados na abordagem I (Tabela 48), contudo a precisão na abordagem II foi maior (Tabela 79).

Tabela 80 – Três melhores resultados de AUC com todos os atributos da abordagem II no problema de novas coautorias.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
meta.Bagging	F	0,72	0,99	0,981	1	1	98,023
	T	0,72	0	0	0	0	
	Avg	0,72	0,971	0,962	0,98	0,981	
trees.RandomForest	F	0,701	0,99	0,981	1	0,996	98,044
	T	0,701	0,007	0,25	0,004	0	
	Avg	0,701	0,971	0,966	0,98	0,977	
meta.RotationForest	F	0,692	0,99	0,981	1	0,996	98,065
	T	0,692	0,007	1	0,004	0	
	Avg	0,692	0,971	0,981	0,981	0,977	

Fonte: William T. Maruyama, 2015.

O maior valor de AUC foi de 0,72, obtido pelo algoritmo *Bagging*, que entre os três ranqueados registrou menor acurácia e classificou todas as instâncias como da classe F (Tabela 80). Quanto à comparação entre o conjunto completo da abordagem I (Tabela 49), a métrica AUC do teste atual também apresentou valores menores.

O algoritmo *Naive Bayes Updateable* foi o que apresentou maior valor de Medida-F (0,114, Tabela 81), sendo ele, como as métricas anteriores, de desempeho inferior à obtida na abordagem I (Tabela 50).

Tabela 81 – Três melhores resultados da Medida-F com todos os atributos da abordagem II no problema de novas coautorias.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
bayes.NaiveBayes Updateable	F	0,677	0,949	0,986	0,914	0,679	90,271
	T	0,677	0,114	0,069	0,321	0,086	
	Avg	0,677	0,932	0,968	0,903	0,667	
bayes.NaiveBayes	F	0,677	0,949	0,986	0,914	0,679	90,271
	T	0,677	0,114	0,069	0,321	0,086	
	Avg	0,677	0,932	0,968	0,903	0,667	
trees.RandomTree	F	0,536	0,972	0,982	0,962	0,891	94,537
	T	0,536	0,072	0,054	0,109	0,038	
	Avg	0,536	0,954	0,964	0,945	0,874	

Fonte: William T. Maruyama, 2015.

5.3.2.2 Abordagem II com todos atributos e balanceamento no problema de novas coautorias

Os resultados dos testes com todos os atributos e com balanceamento do conjunto de dados na abordagem II foram ranqueados conforme a acurácia (Tabela 82), a revocação da classe positiva (Tabela 83), a AUC (Tabela 84) e a Medida-F (Tabela 85).

Tabela 82 – Três melhores resultados de acurácia com todos os atributos da abordagem II no problema de novas coautorias, com balanceamento.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
rules.ZeroR	F	0,5	0,99	0,981	1	1	98,058
	T	0,5	0	0	0	0	
	Avg	0,5	0,971	0,962	0,981	0,981	
meta.Vote	F	0,5	0,99	0,981	1	1	98,058
	T	0,5	0	0	0	0	
	Avg	0,5	0,971	0,962	0,981	0,981	
meta.Stacking	F	0,5	0,99	0,981	1	1	98,058
	T	0,5	0	0	0	0	
	Avg	0,5	0,971	0,962	0,981	0,981	

Fonte: William T. Maruyama, 2015.

Ao balancear os dados na abordagem II, o desempenho da acurácia diminuiu quando comparado com os dados não balanceados (Tabela 82 e 78, respectivamente). O mesmo ocorreu com relação ao balanceamento realizado na abordagem I (Tabela 51), mas os valores do teste atual foram inferiores.

Quanto à revocação, o 1º e o 2º colocados classificaram todas as instâncias como positivas, portanto a precisão foi baixa (Tabela 83). Situação semelhante foi registrada após o balanceamento na abordagem I (Tabela 52). Além disso, o aumento da revocação quando comparado aos resultados do teste não balanceado implicou na diminuição da precisão (Tabela 79).

Tabela 83 – Três melhores resultados de revocação da classe positiva com todos os atributos da abordagem II no problema de novas coautorias, com balanceamento.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
meta.StackingC	F	0,5	0	0	0	0	1,942
	T	0,5	0,038	0,019	1	1	
	Avg	0,5	0,001	0	0,019	0,019	
bayes.DMNBtext	F	0,333	0	0	0	0	1,942
	T	0,333	0,038	0,019	1	1	
	Avg	0,333	0,001	0	0,019	0,019	
rules.ConjunctiveRule	F	0,545	0,347	0,989	0,21	0,12	22,329
	T	0,545	0,042	0,022	0,88	0,79	
	Avg	0,545	0,341	0,97	0,223	0,133	

Fonte: William T. Maruyama, 2015.

Tabela 84 – Três melhores resultados de AUC com todos os atributos da abordagem II no problema de novas coautorias, com balanceamento.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
trees.LADTree	F	0,721	0,905	0,988	0,834	0,522	82,738
	T	0,721	0,097	0,054	0,478	0,166	
	Avg	0,721	0,889	0,97	0,827	0,515	
trees.RandomForest	F	0,693	0,989	0,981	0,998	0,989	97,902
	T	0,693	0,02	0,107	0,011	0,002	
	Avg	0,693	0,971	0,964	0,979	0,97	
lazy.LWL	F	0,692	0,788	0,987	0,656	0,42	65,455
	T	0,692	0,061	0,032	0,58	0,344	
	Avg	0,692	0,774	0,969	0,655	0,418	

Fonte: William T. Maruyama, 2015.

Os valores de AUC foram maiores em relação aos resultados obtidos utilizando-se o conjunto de treinamento não balanceado na abordagem II e ao balanceado na abordagem I (Tabelas 84, 80 e 53, respectivamente).

Tabela 85 – Três melhores resultados da Medida-F com todos os atributos da abordagem II no problema de novas coautorias, com balanceamento.

Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
functions.RBFNetwork	F	0,681	0,938	0,986	0,893	0,624	88,343
	T	0,681	0,111	0,065	0,376	0,107	
	Avg	0,681	0,922	0,968	0,883	0,614	
bayes.NaiveBayes Updateable	F	0,677	0,922	0,987	0,866	0,573	85,721
	T	0,677	0,104	0,059	0,427	0,134	
	Avg	0,677	0,907	0,969	0,857	0,564	
bayes.NaiveBayes	F	0,677	0,922	0,987	0,866	0,573	85,721
	T	0,677	0,104	0,059	0,427	0,134	
	Avg	0,677	0,907	0,969	0,857	0,564	

Fonte: William T. Maruyama, 2015.

Quanto ao ranqueamento da medida-F, é possível notar que os valores de revocação aumentaram em relação ao testes da abordagem I sem balanceamento (Tabela 50) e da abordagem II sem balanceamento (Tabela 81). Porém, os valores de precisão foram menores, logo o resultado da Medida-F da abordagem II balanceada foi inferior (Tabela 85).

5.4 Normalização e PCA do conjunto completo de atributos

Apesar de alguns algoritmos de classificação do Weka realizarem a normalização, foram testados adicionalmente dois métodos de normalização para verificar se há melhoria nas métricas analisadas (ver Seção 2.4.1). O primeiro método mapeou os valores de cada atributo para o intervalo entre [0,1] subtraindo-se de todos os valores o menor encontrado e dividindo o valor resultado pela diferença entre o maior e o menor valores para o respectivo atributo. Para os atributos de distâncias, os valores foram mapeados inversamente, isto é, o maior valor foi mapeado para 0 e o menor para 1. Já o segundo método transformou os dados utilizando uma distribuição em torno da média e do desvio padrão de cada atributo.

Os resultados não apresentaram diferença significativa em relação aos obtidos nos testes sem normalização e, por isso, não são apresentados nesta dissertação.

Também foi testada a utilização da PCA apesar de alguns algoritmos já aplicarem. Com a configuração padrão da PCA no Weka (porcentagem de variância acumulada de 95%) não foram obtidos resultados promissores, portanto não são apresentados nesta dissertação.

5.5 Discussão

No presente trabalho foram realizados experimentos para a predição de relacionamentos de coautorias com diferentes combinações de atributos de domínio e estruturais e foram analisados os resultados obtidos em diferentes experimentos. Tal delineamento foi diferente de alguns trabalhos que utilizaram apenas atributos estruturais (PAVLOV, 2007; LICHTENWALTER; LUSSIER; CHAWLA, 2010; CUKIERSKI; HAMNER; YANG, 2011). Os experimentos realizados utilizaram a estratégia de aprendizagem supervisionado. Conforme observado na literatura, esta estratégia apresenta resultados relevantes com a combinação adequada de atributos (LU et al., 2010; FIRE et al., 2011; SA; PRUDENCIO, 2011; SOARES; PRUDENCIO, 2012; DIGIAMPIETRI; SANTIAGO; ALVES, 2013). Para analisar a relação de custo e benefício dessas variações de conjuntos de atributos nos experimentos, as métricas avaliadas foram a acurácia (taxa de acerto geral), a revocação da classe positiva (ou sensibilidade da classe positiva, que quantifica a fração das instâncias positivas que foram efetivamente classificadas como positivas), a AUC (área sob a curva ROC) e a Medida-F da classe positiva (média harmônica entre a precisão e a revocação da classe positiva).

Levando em consideração que o desbalanceamento de classes pode afetar o desempenho dos algoritmos de classificação (RATTIGAN; JENSEN, 2005; HASAN; ZAKI, 2011), foi utilizada a técnica de *Oversampling* no conjunto de treinamento, em alguns experimentos, para diminuir os efeitos do desbalanceamento e verificar se há melhoria nos resultados. Alguns trabalhos na literatura apresentam bom desempenho na predição, no entanto ignoram a distribuição das classes. Um exemplo é em Hasan et al. (2006), no qual os autores relataram um bom resultado de previsão de *links* em conjuntos de dados da DBLP e da BIOBASE, utilizando-se de atributos estruturais e de domínio. Contudo eles ignoraram a distribuição de classes e utilizaram validação cruzada em um conjunto de dados, no qual a distribuição é equilibrada.

5.5.1 O problema geral de predição de coautorias

Inicialmente, foram realizados testes com o conjunto de dados resultante do filtro horizontal apenas (abordagem I). Apesar do conjunto total de atributos apresentar resultados promissores (ou seja, com acurácia maior que o valor base) no problema geral, com a utilização de algumas estratégias diferentes foi possível obter desempenhos melhores conforme as métricas analisadas. Um exemplo é o experimento de seleção de atributos, que apresentou valores maiores de acurácia do que os obtidos para o conjunto total de atributos. A utilização dos atributos de domínio e dos atributos individuais (no caso, o atributo Conferências presente) também apresentaram resultados superiores. Tais resultados indicam que alguns atributos contribuem mais na discriminação das classes.

Os melhores valores de revocação da classe positiva, por sua vez, foram registrados com o balanceamento no conjunto total de atributos, na seleção de atributos e nos atributos individuais. Neles, os melhores resultados para esta métrica classificaram todas instâncias como da classe “serão coautores” (T), contudo foi registrado uma baixa precisão - e baixa acurácia, consequentemente. Entretanto, ao observar colocações menos altas no ranqueamento desses testes, a taxa de revocação diminui e a acurácia tende a aumentar. Isso porque os classificadores identificam mais classes negativas, contribuindo para o aumento do acerto geral. Desse modo, em determinadas aplicações pode ser interessante escolher uma revocação menor para obter uma acurácia maior.

A métrica AUC mensura a área sob a curva ROC, que pode auxiliar na análise do custo e benefício entre os positivos e os falsos positivos. Nesse contexto, os maiores valores foram registrados no experimento com o conjunto total de atributos.

O maior equilíbrio (Medida-F) entre os valores de precisão e revocação da classe positiva no problema geral foi de 53%. Esse valor foi alcançado no experimento com o subconjunto de atributos de domínio, o Subconjunto 1 da seleção de atributos e o atributo "Conferências presente" (balanceado e não balanceado). Isso indica que o melhor equilíbrio entre as métricas precisão e revocação da classe positiva foi um pouco superior à 50%. Portando, o sistema classificou corretamente aproximadamente metade dos casos positivos da metade dos casos que existem.

Os resultados da seleção de atributos (filtro vertical dos dados) mostraram que o Subconjunto 1 - composto pelos atributos Periódicos presente, Conferências presente, Distância geográfica e Distância no grafo (Tabela 24) – apresentou os valores de taxa de acerto, de revocação da classe positiva e de Medida-F maiores que os apresentados no conjunto total de atributos, mas os valores de AUC foram menores. Contudo, o conjunto de atributos do Subconjunto 1 foi bastante reduzido por utilizar apenas quatro atributos (2 de domínio/contexto e 2 estruturais), diminuindo a dimensão do problema. Isso pode indicar que alguns atributos não estavam contribuindo para classificação e poderiam estar tornando mais complexa a classificação com o aumento da dimensão do problema. Outro fato a se notar na formação dos subconjuntos do teste de seleção é que todos os algoritmos consideraram os atributos Periódicos presente e Conferências presente como relevantes. Na matriz de correlação (Figura 11), por sua vez, os mesmos atributos foram dois dos que obtiveram maiores valores de correlação com a classe, pois aumentam as chances de futuras publicações por serem informações de parcerias recentes entre autores.

Liben-Nowell e Kleinberg (2003) apresentaram resultados promissores com o atributo *Katz*, assim como os obtidos no presente trabalho, contudo os autores mediram o desempenho em relação a um método aleatório de predição. Já o atributo Distância no grafo no experimento dos mesmos autores, de modo geral, não apresentou resultados melhores que os outros atributos. Diferentemente, Hasan et al. (2006) realizaram um ranqueamento dos atributos, sendo que o atributo Distância no grafo (*Shortest Path* ou SP) foi um dos melhores colocados nos dois conjuntos de dados analisados, assim como no ranqueamento realizado no presente estudo.

Para verificar a influência nos resultados das instâncias positivas (serão coautores) que foram eliminadas do conjunto de treinamento pelo filtro horizontal, foi realizada a abordagem II. Em relação ao experimento da abordagem I com todos os atributos e sem balanceamento, os valores de acurácia foram muito semelhantes. Mas para as outras métricas (revocação, AUC e Medida-F), os resultados foram próximos e com pequena vantagem para os obtidos na abordagem I. No geral, entre as duas abordagens testadas, não foi observada uma diferença de valores discrepante entre as métricas analisadas. Isto é, a diminuição da quantidades de dados favoreceu a eficiência no processo de treinamento, não eliminou muitas instâncias positivas e os resultados não foram prejudicados. Desse modo, pode-se considerar que a utilização da metodologia e seus critérios no filtro horizontal foi eficiente no problema de predição geral.

5.5.2 O problema de predição de novas coautorias

No problema de novas coautorias evidenciou-se mais evidente o desbalanceamento de classes, pois o número de instâncias da classe positiva era significativamente menor. Isso ocorre porque nesse problema se lida com parcerias novas - ou seja, que não existem na janela de tempo definida como presente -, as quais são casos menos frequentes em uma rede acadêmica. Portanto, existem menos exemplos de casos positivos nos dados e os classificadores tendem a classificar tudo como negativo. Logo, devido à complexidade do problema, nenhum classificador foi capaz de alcançar uma acurácia muito acima do valor definido como base no presente trabalho.

Os maiores valores de revocação da classe positiva, na abordagem I, foram registrados nos experimentos com balanceamento, no qual os primeiros colocados classificaram todas instâncias como positiva e, portanto, registrando baixa precisão e acurácia. Excluindo os testes com o conjunto de treinamento balanceado, o maior valor de revocação em novas coautorias foi registrado pelos três atributos individuais ranqueados (PA, Artigos em conferência 2 e Periódicos presente) em que todas as instâncias foram classificadas como positivas, também apresentando valores muito baixos de precisão e acurácia. Em segundo lugar está o experimento com a utilização de filtros verticais, que também apresentou os maiores valores de AUC (Subconjunto 6) e de Medida-F (Subconjunto 1) registrados nos experimentos. Isso pode indicar que a utilização de um pré-processamento no conjunto de atributos, formando subconjuntos a partir de uma filtragem, pode atenuar parte da

complexidade do respectivo problema e favorecer a classificação de mais casos verdadeiros positivos do que quando com o conjunto completo de atributos.

O atributo Conferência anterior foi o mais relevante segundo os algoritmos de seleção de atributos, pois esteve presente em todos os subconjuntos formados por esses algoritmos. No ranqueamento dos atributos, por sua vez, ele está em terceiro lugar (mediana 5). Diferentemente do problema geral, o atributo Distância no grafo não foi um dos melhores colocados, ficando em oitavo lugar no ranqueamento dos atributos (mediana 12). Contudo, em uma análise geral do teste de atributos individuais, nenhum atributo que se destacou nas métricas e conforme a matriz de correlação (veja a Figura 12) teve um desempenho muito bom na predição.

Os resultados não se apresentaram muito diferentes entre as duas abordagens testadas, excetuando que a revocação de verdadeiros positivos foi um pouco maior na abordagem I. Do mesmo modo que no problema geral, pode-se considerar que a utilização do filtro horizontal foi eficiente no atual problema de coautorias inéditas.

5.6 Considerações Finais

O presente trabalho foi delineado com o intuito de desenvolver uma solução para a predição de relacionamentos de coautoria que considere a combinação de diferentes atributos e filtros. Para tal, a predição de relacionamento foi dividida em dois problemas: problema geral e de novas coautorias/inéditas.

Os problemas de predição foram tratados como um problema de classificação, no qual foram extraídos/calculados 30 atributos. Sendo desse conjunto, 15 atributos estruturais e 15 de domínio/contexto da rede social em estudo. Diferentes (sub)conjuntos foram formados pela combinação dos 30 atributos e eles foram submetidos a diferentes algoritmos disponíveis no Weka. Em dois subconjuntos foi realizado o balanceamento do conjunto de treinamento para verificar se há ou não melhora no desempenho dos resultados. Resultados estes que foram avaliados por diferentes métricas (acurácia, revocação da classe positiva, AUC e Medida-F) para analisar diferentes aspectos dos resultados.

No problema geral, podemos observar resultados acima do valor base com a seleção de atributos (filtro vertical), atributos de domínio, conjunto total de atributos e dos atributos individuais. Com o conjunto total também foi possível obter valores mais altos em AUC. Os maiores valores da Medida-F foram alcançados com um subconjunto de atributos

de domínio, o Subconjunto 1 da seleção de atributos e o atributo Conferências presente (balanceado e não balanceado). Já no problema de novas coautorias, os classificadores obtiveram apenas valores próximos ao valor definido como base. O maior valor de AUC foi alcançado com a seleção de características (Subconjunto 6) e a maior Medida-F também (com o Subconjunto 1).

Por meio dos experimentos com o balanceamento nos dois problemas, foi possível alcançar valores altos de revocação da classe positiva. Contudo, o valor foi extremo (isto é, 1 de revocação) nas primeiras colocações, classificando todas as instâncias como positivas.

Tabela 86 – Os 4os colocados no ranqueamento de revocação da classe positiva, sem balanceamento.

Problema	Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
Geral	bayes.NaiveBayes	F	0,874	0,943	0,985	0,905	0,28	89,601
		T	0,87	0,397	0,274	0,72	0,095	
		Avg	0,873	0,917	0,951	0,896	0,271	
Novas Coautorias	bayes.NaiveBayes Updateable	F	0,742	0,949	0,986	0,914	0,675	90,271
		T	0,742	0,115	0,07	0,325	0,086	
		Avg	0,742	0,932	0,968	0,903	0,664	

Fonte: William T. Maruyama, 2015.

Tabela 87 – Os 5os colocados no ranqueamento de revocação da classe positiva, com balanceamento.

Problema	Classificador		AUC	Medida-F	Precisão	Revocação	taxa FP	Acurácia (%)
Geral Balanceado	bayes.BayesNet	F	0,88	0,9	0,987	0,826	0,215	82,424
		T	0,88	0,298	0,184	0,785	0,174	
		Avg	0,88	0,871	0,949	0,824	0,213	
Novas Coautorias Balanceado	trees.LADTree	F	0,72	0,768	0,99	0,628	0,336	62,826
		T	0,72	0,065	0,034	0,664	0,372	
		Avg	0,72	0,754	0,971	0,628	0,336	

Fonte: William T. Maruyama, 2015.

As Tabelas 86 e 87 apresentam resultados que estão os 4º e 5º colocados, respectivamente, o ranqueamento da revocação da classe positiva e é possível observar valores de acurácia maiores ao diminuir a revocação. Portanto, é necessário analisar o custo e beneficio de um revocação mais baixa, porém com um acurácia mais alta.

Além disso, os resultados entre as duas abordagens demonstram que o filtro horizontal foi eficiente, ao diminuir a dimensão da rede, isto é, o número de instâncias a serem analisadas. Isso porque a inclusão no conjunto de dados das instâncias positivas eliminadas pelo filtro, não apresentou uma diferença discrepante dos resultados entre ambas abordagens.

Como são apresentados pelos resultados do presente trabalho, as combinações de atributos e os pré-processamentos realizados propiciaram diferentes valores nas métricas avaliadas, tanto no problema geral quanto no de novas coautorias. Contudo, a partir

destes resultados não foi possível encontrar uma única/melhor solução para a predição de relacionamentos de coautoria, mostrando que se faz necessário analisar o custo-benefício de cada métrica sobre cada experimento realizado. Isto é, cada métrica apresentou o melhor desempenho em experimentos diferentes.

6 Conclusões e Trabalhos Futuros

Como mencionado em capítulos anteriores, a predição de *links* em redes sociais é uma atividade complexa, com diversos desafios.

Para o presente projeto, o primeiro desafio foi identificar quais atributos estão sendo utilizados em outros trabalhos de redes sociais e quais poderiam ser utilizados no contexto deste projeto. Deste modo, uma revisão sistemática foi inicialmente realizada (Capítulo 3) e indicou, em linhas gerais, uma ampla abrangência de quais atributos e metodologias estão sendo utilizadas no estado da arte. Dentre os vários atributos identificados, foram selecionados alguns atributos estruturais amplamente utilizados na literatura e acrescentados outros de domínio da rede que foram identificados pelos autores do presente trabalho.

A rede social acadêmica estudada foi elaborada a partir das informações acadêmicas extraídas da Plataforma Lattes. Ela é formada por pesquisadores permanentes dos programas de pós-graduação em Ciências da Computação (detalhes no Capítulo 4). Nesta rede, a predição de relacionamentos foi tratada como dois problemas diferentes: problema geral (predição de *links*) e problema de novas coautorias (predição de *links* inéditos).

Neste contexto - e considerando os desafios mencionados como a combinação de atributos e o fato dos conjuntos de dados serem tipicamente desbalanceados (Capítulo 1) - foram montados diferentes (sub)conjuntos de dados e eles foram avaliados a partir de quatro métricas. Com exceção dos subconjuntos de domínio e estrutural, foram aplicadas técnicas de seleção de características, com intuito de encontrar subconjuntos com atributos relevantes segundo elas. Para o balanceamento dos dados foi aplicada a técnica de *Oversampling* no conjunto de treinamento. Além disso, para verificar a influência do filtro horizontal, testou-se uma abordagem diferente, a qual incluia no conjunto de treinamento as instâncias da classe positiva excluídas pelo filtro (abordagem II).

Observa-se que alguns atributos se destacaram de acordo com os resultados obtidos com a seleção de atributos no problema geral. Essa seleção, além de ter obtido maior acurácia, pode ser vantajosa na diminuição da quantidade de dados e do tempo de processamento dos algoritmos na fase de treinamento. Apesar da combinação com a seleção de atributos alcançar maior acurácia, a com todos os atributos registrou maior AUC, isto é, teve maior equilíbrio entre os verdadeiros positivos e o falsos positivos.

Nos dois problemas, os atributos (*Coautorias em*) *Conferências* "presente" no problema geral e "passado" no problema de novas coautorias foram importantes nos experimentos de classificação, já que eles obtiveram uma boa colocação no ranqueamento dos atributos individuais. Considerando que estes atributos estão relacionados à publicação em coautoria em conferências (tipo de publicação muito importante na área de ciência da computação) é natural que estes atributos tenham se destacados. Dentro deste contexto, a criação de um novo atributo indicando a frequência em que dois pesquisadores possuem artigos publicados em um mesmo evento científico (e/ou participaram de um mesmo evento) poderia, potencialmente, auxiliar no processo de predição.

Quanto ao uso da técnica de balanceamento de dados, ela apresentou melhor desempenho na recomendação da classe positiva (maior revocação), mas redução da precisão devido ao aumento de falsos positivos nos dois problemas estudados. Este comportamento é esperado e, dependendo do tipo de aplicação, pode-se preferir uma maior precisão ou uma maior revocação. Para problemas com esta complexidade, se faz necessário avaliar o custo-benefício das diferentes abordagens (e de classificadores) em relação ao objetivo da predição (por exemplo, maximizar a precisão da classe positiva ou maximizar a revocação da classe positiva tendo uma acurácia global "satisfatória"). Além disso, a realização das duas abordagens demonstrou que, devido aos valores próximos (ou inferiores) obtidos nas métricas, a utilização do filtro horizontal foi eficiente. Isso porque a inclusão das instâncias positivas descartadas pelo uso do respectivo filtro (abordagem II) não apresentou valores das métricas muito discrepantes das filtradas (abordagem I), indicando que a diminuição do volume de dados apresentou resultados satisfatórios no processo de treinamento.

A partir dos resultados registrados, pode-se observar que o desempenho de cada métrica variou conforme a estratégia utilizada e que não houve uma estratégia que apresentasse todas as métricas com o melhor desempenho. Isso porque há uma relação de custo-benefício de cada estratégia de combinação dos atributos e de técnicas de pré-processamento utilizados. Portanto, para cada métrica houve um experimento que retornou uma melhor solução para cada problema - geral ou novas coautorias.

Ao avaliar o custo benefício entre a revocação da classe positiva e a acurácia, o erro também pode ser tolerável em certa medida, pois ao recomendar/classificar a ocorrência de uma coautoria, mesmo que seja um erro (segundo a classe), pode-se incentivar que ocorra uma colaboração entre os pesquisadores. Isto é, podem ocorrer casos em que a solução classificou como positiva, pois os valores dos atributos indicam que ocorreria uma

parceria, contudo não ocorreu na janela de tempo testada. Logo, mesmo quando o erro de classificação ocorre, a indicação pode ser interessante para favorecer a comunicação entre os pesquisadores e incentivar uma parceria.

6.1 Principais Contribuições

O presente projeto teve como objetivo principal uma solução que realizou diferentes combinações de atributos e filtros. Para atingi-lo, foi adotado a estratégia de aprendizagem supervisionada e analisou-se diferentes (sub)conjuntos de atributos considerando os resultados de quatro métricas (acurácia, revocação de classe positiva, AUC e Medida-F). O presente trabalho também teve as seguintes contribuições:

- Uma revisão sistemática sobre a Predição de *Links* em redes sociais. Esta revisão pode ser útil para identificar os atributos e métodos utilizados na predição de relacionamentos e também pode servir como base para futuros trabalhos;
- A identificação de atributos para aplicação em análise de redes de coautorias. Principalmente na identificação dos atributos de domínio, pois são atributos específicos do contexto da aplicação;
- Disponibilização do conjunto de dados utilizados nos experimentos;
- Desenvolvimento de uma solução que automatiza o processo da predição de coautorias, utilizado neste trabalho. Esta solução pode auxiliar futuros experimentos e futuras extensões;
- Parte dos resultados desta dissertação foram utilizados na publicação de artigos científicos (Digiampietri e Maruyama (2014), Digiampietri et al. (2015)).

6.2 Trabalhos Futuros

Algumas possibilidades de continuidade e melhoria do projeto (tanto para o problema geral quanto para novas coautorias) são apresentadas a seguir:

- Mais caracterização dos dados utilizados;
- Experimentos com estratégias de balanceamento diferentes;
- Experimentos com variações nos parâmetros dos algoritmos de classificação;

- Uso de atributos adicionais como *Participação em Eventos em Comum* (e/ou *Frequência de publicação em um mesmo evento*) e *Participação em Bancas em Comum*;
- Extração e combinação de atributos oriundos de informações de outras fontes de informação;
- Experimentos com relacionamentos ponderados;
- Experimentos considerando atributos temporais (e/ou utilizando diferentes janelas de tempo para o cálculo dos atributos).
- Experimentos considerando atributos temporais (e/ou utilizando diferentes janelas de tempo para o cálculo dos atributos).
- Abordar como um problema multiclasse. Na qual, poderiamos ter os seguintes possíveis rótulos:
  - Não colaboravam e não colaborarão;
  - Não colaboravam e colaborarão;
  - Colaboravam e não colaboravam;
  - Colaboravam e colaborarão.

Referências[1]

ADAMIC, L.; ADAR, E. Friends and neighbors on the web. *Social Networks*, v. 25, p. 211–230, 2001. Citado na página 39.

AHA, D. W.; KIBLER, D.; ALBERT, M. K. Instance-based learning algorithms. *Mach. Learn.*, Kluwer Academic Publishers, Hingham, MA, USA, v. 6, n. 1, p. 37–66, jan. 1991. ISSN 0885-6125. Disponível em: <http://dx.doi.org/10.1023/A:1022689900470>. Citado na página 51.

AIELLO, L. M. et al. Friendship prediction and homophily in social media. *ACM Trans. Web*, ACM, New York, NY, USA, v. 6, n. 2, p. 9:1–9:33, jun. 2012. ISSN 1559-1131. Citado 3 vezes nas páginas 62, 73 e 74.

ALMANSOORI, W. et al. Link prediction and classification in social networks and its application in healthcare. In: *Information Reuse and Integration (IRI), 2011 IEEE International Conference on*. [S.l.: s.n.], 2011. p. 422–428. Citado na página 62.

ATKESON, C. G.; MOORE, A. W.; SCHAAL, S. Locally weighted learning. *Artif. Intell. Rev.*, Kluwer Academic Publishers, Norwell, MA, USA, v. 11, n. 1-5, p. 11–73, fev. 1997. ISSN 0269-2821. Disponível em: <http://dx.doi.org/10.1023/A:1006559212014>. Citado na página 51.

BARABáSI, A. L. et al. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, v. 311, n. 3-4, p. 590–614, 2002. ISSN 03784371. Citado na página 39.

BIOLCHINI, J. et al. *Systematic Review in Software Engineering*. Rio de Janeiro, 2005. 30 p. Citado na página 60.

BISHOP, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006. ISBN 0387310738. Citado na página 45.

BOUCKAER, R. R. *Bayesian Network Classifiers in Weka for Version 3-5-7*. [S.l.], 2008. Citado na página 49.

BRADLEY, A. P. The use of the area under the {ROC} curve in the evaluation of machine learning algorithms. *Pattern Recognition*, v. 30, n. 7, p. 1145 – 1159, 1997. ISSN 0031-3203. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0031320396001422>. Citado na página 57.

BREIMAN, L. Bagging predictors. *Machine Learning*, v. 24, n. 2, p. 123–140, 1996. Citado na página 56.

BREIMAN, L. Random forests. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001. Citado na página 54.

BREIMAN, L. et al. *Classification and Regression Trees*. [S.l.]: Taylor & Francis, 1984. ISBN 9780412048418. Citado na página 53.

[1] De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

BUCKLAND, M.; GEY, F. The relationship between recall and precision. *J. Am. Soc. Inf. Sci.*, John Wiley & Sons, Inc., New York, NY, USA, v. 45, n. 1, p. 12–19, jan. 1994. ISSN 0002-8231. Disponível em: <http://dx.doi.org/10.1002-/(SICI)1097-4571(199401)45:1¡12::AID-ASI2¿3.0.CO;2-L>. Citado na página 59.

CESSIE, S. le; HOUWELINGEN, J. van. Ridge estimators in logistic regression. *Applied Statistics*, v. 41, n. 1, p. 191–201, 1992. Citado na página 51.

CHANG, C.; YAO, X. Social network link predict based on af model. In: *Computer Science and Network Technology (ICCSNT), 2011 International Conference on*. [S.l.: s.n.], 2011. v. 1, p. 415–418. Citado 2 vezes nas páginas 15 e 62.

CHELMIS, C.; PRASANNA, V. Predicting communication intention in social networks. In: *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. [S.l.: s.n.], 2012. p. 184–194. Citado na página 62.

CLEARY, J. G.; TRIGG, L. E. K*: An instance-based learner using an entropic distance measure. In: *12th International Conference on Machine Learning*. [S.l.: s.n.], 1995. p. 108–114. Citado na página 51.

COHEN, W. W. Fast effective rule induction. In: *Twelfth International Conference on Machine Learning*. [S.l.]: Morgan Kaufmann, 1995. p. 115–123. Citado na página 52.

CORLETTE, D.; SHIPMAN III, F. M. Link prediction applied to an open large-scale online social network. In: *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*. New York, NY, USA: ACM, 2010. (HT '10), p. 135–140. ISBN 978-1-4503-0041-4. Citado 2 vezes nas páginas 62 e 77.

CORMEN, T.; LEISERSON, C.; STEIN, R. *ALGORITMOS*. [S.l.]: CAMPUS - RJ, 2012. ISBN 8535236996. Citado 3 vezes nas páginas 31, 32 e 33.

COSTA, G.; ORTALE, R. A bayesian hierarchical approach for exploratory analysis of communities and roles in social networks. In: *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*. [S.l.: s.n.], 2012. p. 194–201. Citado na página 62.

CUKIERSKI, W.; HAMNER, B.; YANG, B. Graph-based features for supervised link prediction. In: *Neural Networks (IJCNN), The 2011 International Joint Conference on*. [S.l.: s.n.], 2011. p. 1237–1244. ISSN 2161-4393. Citado 3 vezes nas páginas 28, 62 e 134.

DASH, M.; LIU, H. Feature selection for classification. *Intelligent Data Analysis*, v. 1, n. 14, p. 131 – 156, 1997. ISSN 1088-467X. Disponível em: <http://www-.sciencedirect.com/science/article/pii/S1088467X97000085>. Citado na página 46.

DIGIAMPIETRI, L.; MARUYAMA, W. Predição de novas coautorias na rede social acadêmica dos programas brasileiros de pós-graduação em ciência da computação. In: *CSBC 2014 - BraSNAM*. [S.l.: s.n.], 2014. Citado na página 143.

DIGIAMPIETRI, L. et al. Minerando e caracterizando dados de currículos lattes. In: *CSBC 2012 - BraSNAM*. [S.l.: s.n.], 2012. Citado na página 83.

DIGIAMPIETRI, L. et al. Dinâmica das relações de coautoria nos programas de pós-graduação em computação no brasil. In: *CSBC 2012 - BraSNAM*. [S.l.: s.n.], 2012. Citado 2 vezes nas páginas 83 e 84.

DIGIAMPIETRI, L.; SANTIAGO, C.; ALVES, C. Predição de coautorias em redes sociais acadêmicas: um estudo exploratório em ciência da computação. In: *CSBC 2013 - BraSNAM*. [S.l.: s.n.], 2013. Citado 4 vezes nas páginas 27, 29, 82 e 134.

DIGIAMPIETRI, L. A. et al. Um sistema de predição de relacionamentos em redes sociais. In: *XI Simpósio Brasileiro de Sistemas de Informação (SBSI 2015)*. [S.l.: s.n.], 2015. p. 139–146. Citado na página 143.

DONG, Y. et al. Predicting missing links via local feature of common neighbors. In: *Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth International Conference on*. [S.l.: s.n.], 2011. v. 2, p. 1038–1042. Citado 3 vezes nas páginas 27, 63 e 74.

DONG, Y. et al. Random walk based resource allocation: Predicting and recommending links in cross-operator mobile communication networks. In: *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*. [S.l.: s.n.], 2011. p. 358–365. Citado 2 vezes nas páginas 63 e 76.

DONG, Y. et al. Link prediction and recommendation across heterogeneous social networks. In: *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. [S.l.: s.n.], 2012. p. 181–190. ISSN 1550-4786. Citado 5 vezes nas páginas 26, 63, 72, 73 e 80.

DUDA, R. O.; HART, P. E. *Pattern classification and scene analysis*. New York, London: J. Wiley & Sons, 1973. A Wiley-interscience publication. ISBN 0-471-22361-1. Disponível em: <http://opac.inria.fr/record=b1102308>. Citado na página 49.

FAWCETT, T. An introduction to roc analysis. *Pattern Recogn. Lett.*, Elsevier Science Inc., New York, NY, USA, v. 27, n. 8, p. 861–874, jun. 2006. ISSN 0167-8655. Disponível em: <http://dx.doi.org/10.1016/j.patrec.2005.10.010>. Citado na página 57.

FIRE, M. et al. Link prediction in social networks using computationally efficient topological features. In: *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*. [S.l.: s.n.], 2011. p. 73–80. Citado 3 vezes nas páginas 26, 63 e 134.

FRANK, E.; HALL, M.; PFAHRINGER, B. Locally weighted naive bayes. In: *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2003. (UAI'03), p. 249–256. ISBN 0-127-05664-5. Disponível em: <http://dl.acm.org/citation.cfm?id=2100584.2100614>. Citado na página 51.

FRANK, E. et al. Using model trees for classification. *Machine Learning*, v. 32, n. 1, p. 63–76, 1998. Citado na página 56.

FREUND, Y.; SCHAPIRE, R. E. Experiments with a new boosting algorithm. In: *Thirteenth International Conference on Machine Learning*. San Francisco: Morgan Kaufmann, 1996. p. 148–156. Citado na página 56.

FREUND, Y.; SCHAPIRE, R. E. Large margin classification using the perceptron algorithm. *Mach. Learn.*, Kluwer Academic Publishers, Hingham, MA, USA, v. 37, n. 3, p. 277–296, dez. 1999. ISSN 0885-6125. Disponível em: <http://dx.doi.org/10.1023-/A:1007662407062>. Citado na página 50.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. *Additive Logistic Regression: a Statistical View of Boosting*. Stanford University, 1998. Citado na página 56.

GAMA, J. Functional trees. v. 55, n. 3, p. 219–250, 2004. Citado na página 54.

GAO, S.; DENOYER, L.; GALLINARI, P. Temporal link prediction by integrating content and structure information. In: *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2011. (CIKM '11), p. 1169–1174. ISBN 978-1-4503-0717-8. Citado 4 vezes nas páginas 28, 63, 70 e 80.

GAO, S.; DENOYER, L.; GALLINARI, P. Link prediction via latent factor blockmodel. In: *Proceedings of the 21st International Conference Companion on World Wide Web*. New York, NY, USA: ACM, 2012. (WWW '12 Companion), p. 507–508. ISBN 978-1-4503-1230-1. Citado 2 vezes nas páginas 27 e 63.

GENKIN, A.; LEWIS, D. D.; MADIGAN, D. Large-scale bayesian logistic regression for text categorization. *Technometrics*, v. 49, p. 291–304(14), August 2007. Disponível em: <http://www.ingentaconnect.com/content/asa/tech/2007/00000049/00000003-/art00007>. Citado na página 49.

GETOOR, L.; DIEHL, C. P. Link mining: A survey. *SIGKDD Explor. Newsl.*, ACM, New York, NY, USA, v. 7, n. 2, p. 3–12, dez. 2005. ISSN 1931-0145. Citado na página 44.

GIRVAN, M.; NEWMAN, M. E. J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, v. 99, n. 12, p. 7821–7826, 2002. Disponível em: <http://www.pnas.org/content/99/12/7821.abstract>. Citado na página 39.

GUO, J.; GUO, H. Multi-features link prediction based on matrix. In: *Computer Design and Applications (ICCDA), 2010 International Conference on*. [S.l.: s.n.], 2010. v. 1, p. V1–357–V1–361. Citado 3 vezes nas páginas 27, 64 e 75.

HALL, M.; FRANK, E. Combining naive bayes and decision tables. In: *Proceedings of the 21st Florida Artificial Intelligence Society Conference (FLAIRS)*. [S.l.]: AAAI press, 2008. p. 318–319. Citado na página 52.

HALL, M. et al. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, ACM, New York, NY, USA, v. 11, n. 1, p. 10–18, nov. 2009. ISSN 1931-0145. Citado na página 85.

HAN, J.; KAMBER, M. *Data Mining: Concepts and Techniques*. [S.l.]: Elsevier, 2012. ISBN 9789380931913. Citado 2 vezes nas páginas 44 e 45.

HASAN, M.; ZAKI, M. A survey of link prediction in social networks. In: AGGARWAL, C. C. (Ed.). *Social Network Data Analytics*. [S.l.]: Springer US, 2011. p. 243–275. ISBN 978-1-4419-8461-6. Citado 9 vezes nas páginas 26, 28, 36, 38, 39, 40, 41, 42 e 135.

HASAN, M. A. et al. Link prediction using supervised learning. In: *In Proc. of SDM 06 workshop on Link Analysis, Counterterrorism and Security*. [S.l.: s.n.], 2006. Citado 5 vezes nas páginas 36, 41, 42, 135 e 136.

HOLMES, G. et al. Multiclass alternating decision trees. In: *ECML*. [S.l.]: Springer, 2001. p. 161–172. Citado na página 54.

HSIEH, C.-J. et al. Organizational overlap on social networks and its applications. In: *Proceedings of the 22Nd International Conference on World Wide Web*. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2013. (WWW '13), p. 571–582. ISBN 978-1-4503-2035-1. Citado 3 vezes nas páginas 26, 64 e 79.

HUANG, J.; LING, C. Using auc and accuracy in evaluating learning algorithms. *Knowledge and Data Engineering, IEEE Transactions on*, v. 17, n. 3, p. 299–310, March 2005. ISSN 1041-4347. Citado na página 58.

HUANG, J. et al. Trust prediction via aggregating heterogeneous social networks. In: *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2012. (CIKM '12), p. 1774–1778. ISBN 978-1-4503-1156-4. Citado 4 vezes nas páginas 15, 64, 72 e 73.

JACCARD, P. *Etude comparative de la distribution florale dans une portion des Alpes et du Jura*. [S.l.]: Impr. Corbaz, 1901. Citado na página 38.

JAMALI, M.; HUANG, T.; ESTER, M. A generalized stochastic block model for recommendation in social rating networks. In: *Proceedings of the Fifth ACM Conference on Recommender Systems*. New York, NY, USA: ACM, 2011. (RecSys '11), p. 53–60. ISBN 978-1-4503-0683-6. Citado na página 64.

JOHN, G. H.; LANGLEY, P. Estimating continuous distributions in bayesian classifiers. In: *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995. (UAI'95), p. 338–345. ISBN 1-55860-385-9. Disponível em: <http://dl.acm.org/citation.cfm?id=2074158.2074196>. Citado na página 49.

KAMEI, T. et al. Predicting missing links in social networks with hierarchical dirichlet processes. In: *Neural Networks (IJCNN), The 2012 International Joint Conference on*. [S.l.: s.n.], 2012. p. 1–8. ISSN 2161-4393. Citado na página 64.

KATZ, L. A new status index derived from sociometric analysis. *Psychometrika*, Springer-Verlag, v. 18, n. 1, p. 39–43, 1953. ISSN 0033-3123. Disponível em: <http://dx.doi.org/10.1007/BF02289026>. Citado na página 40.

KITTLER, J. et al. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 20, n. 3, p. 226–239, 1998. Citado na página 55.

KOHAVI, R. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In: *Second International Conference on Knoledge Discovery and Data Mining*. [S.l.: s.n.], 1996. p. 202–207. Citado na página 54.

KUNCHEVA, L. I. *Combining Pattern Classifiers: Methods and Algorithms*. [S.l.]: John Wiley and Sons, Inc., 2004. Citado na página 55.

KUNEGIS, J.; PREUSSE, J.; SCHWAGEREIT, F. What is the added value of negative links in online social networks? In: *Proceedings of the 22Nd International Conference on World Wide Web*. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2013. (WWW '13), p. 727–736. ISBN 978-1-4503-2035-1. Citado 2 vezes nas páginas 64 e 78.

KUO, T.-T. et al. Unsupervised link prediction using aggregative statistics on heterogeneous social networks. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2013. (KDD '13), p. 775–783. ISBN 978-1-4503-2174-7. Citado 5 vezes nas páginas 20, 28, 64, 72 e 73.

LANDWEHR, N.; HALL, M.; FRANK, E. Logistic model trees. v. 95, n. 1-2, p. 161–205, 2005. Citado 2 vezes nas páginas 50 e 54.

LANGVILLE, A.; MEYER, C. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. [S.l.]: Princeton University Press, 2009. ISBN 9780691122021. Citado na página 41.

LEICHT, E. A.; HOLME, P.; NEWMAN, M. E. J. Vertex similarity in networks. *Phys. Rev. E*, American Physical Society, v. 73, p. 026120, Feb 2006. Citado na página 39.

LERMAN, K. et al. Using proximity to predict activity in social networks. In: *Proceedings of the 21st International Conference Companion on World Wide Web*. New York, NY, USA: ACM, 2012. (WWW '12 Companion), p. 555–556. ISBN 978-1-4503-1230-1. Citado 2 vezes nas páginas 64 e 73.

LEROY, V.; CAMBAZOGLU, B. B.; BONCHI, F. Cold start link prediction. In: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2010. (KDD '10), p. 393–402. ISBN 978-1-4503-0055-1. Citado 2 vezes nas páginas 64 e 79.

LIBEN-NOWELL, D.; KLEINBERG, J. The link prediction problem for social networks. In: *Proceedings of the Twelfth International Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2003. (CIKM '03), p. 556–559. ISBN 1-58113-723-0. Citado 7 vezes nas páginas 26, 27, 38, 40, 64, 70 e 136.

LICHTENWALTER, R. N.; LUSSIER, J. T.; CHAWLA, N. V. New perspectives and methods in link prediction. In: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2010. (KDD '10), p. 243–252. ISBN 978-1-4503-0055-1. Disponível em: <http://doi.acm.org/10.1145/1835804.1835837>. Citado na página 134.

LIN, Z.; YUN, X.; ZHU, Y. Link prediction using benefitranks in weighted networks. In: *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*. Washington, DC, USA: IEEE Computer Society, 2012. (WI-IAT '12), p. 423–430. ISBN 978-0-7695-4880-7. Citado 5 vezes nas páginas 27, 43, 65, 75 e 80.

LIU, X. et al. Co-authorship networks in the digital library research community. *Inf. Process. Manage.*, Pergamon Press, Inc., Tarrytown, NY, USA, v. 41, n. 6, p. 1462–1480, dez. 2005. ISSN 0306-4573. Citado 3 vezes nas páginas 31, 32 e 38.

LÜ, L.; ZHOU, T. Link prediction in complex networks: A survey. *Physica A*, abs/1010.0725, n. 6, p. 1150–1170, 2010. Citado 3 vezes nas páginas 38, 40 e 43.

LU, Z. et al. Supervised link prediction using multiple sources. In: *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. [S.l.: s.n.], 2010. p. 923–928. ISSN 1550-4786. Citado 4 vezes nas páginas 36, 65, 72 e 134.

MAKREHCHI, M. Social link recommendation by learning hidden topics. In: *Proceedings of the Fifth ACM Conference on Recommender Systems*. New York, NY, USA: ACM, 2011. (RecSys '11), p. 189–196. ISBN 978-1-4503-0683-6. Citado 3 vezes nas páginas 27, 65 e 72.

MUGNAINI, R. et al. Normalização de nomes de autores em fontes de informação institucionais: proposta de um método automático de verificação de erros. *Em Questão*, v. 18, n. 3, p. 263–279, 2012. Citado na página 84.

NEWMAN, M. *Networks: An Introduction*. New York, NY, USA: Oxford University Press, Inc., 2010. ISBN 0199206651, 9780199206650. Citado 5 vezes nas páginas 27, 31, 32, 33 e 34.

NEWMAN, M. E. J. Clustering and preferential attachment in growing networks. *Phys. Rev. E*, 2001. Citado na página 39.

NIE, F. et al. Robust matrix completion via joint schatten p-norm and lp-norm minimization. In: *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. [S.l.: s.n.], 2012. p. 566–574. ISSN 1550-4786. Citado na página 65.

OU, Q. et al. Power-law strength-degree correlation from resource-allocation dynamics on weighted networks. *Physical Review E*, v. 75, n. 2, p. 021102, 2007. Citado na página 40.

PAVLOV, M. Finding experts by link prediction in co-authorship networks. *CEUR Workshop Proceedings*, v. 290, p. 42–55, 2007. ISSN 16130073. Citado na página 134.

PEREZ, C.; BIRREGAH, B.; LEMERCIER, M. The multi-layer imbrication for data leakage prevention from mobile devices. In: *Trust, Security and Privacy in Computing and Communications (TrustCom), 2012 IEEE 11th International Conference on*. [S.l.: s.n.], 2012. p. 813–819. Citado 3 vezes nas páginas 26, 65 e 76.

PLATT, J. C. Advances in kernel methods. In: SCHöLKOPF, B.; BURGES, C. J. C.; SMOLA, A. J. (Ed.). Cambridge, MA, USA: MIT Press, 1999. cap. Fast Training of Support Vector Machines Using Sequential Minimal Optimization, p. 185–208. ISBN 0-262-19416-3. Disponível em: <http://dl.acm.org/citation.cfm?id=299094.299105>. Citado na página 50.

PRELL, C. *Social Network Analysis: History, Theory and Methodology*. [S.l.]: SAGE Publications, 2011. ISBN 9781446254103. Citado 3 vezes nas páginas 31, 32 e 34.

QUERCIA, D.; CAPRA, L. Friendsensing: Recommending friends using mobile phones. In: *Proceedings of the Third ACM Conference on Recommender Systems*. New York, NY, USA: ACM, 2009. (RecSys '09), p. 273–276. ISBN 978-1-60558-435-5. Citado 4 vezes nas páginas 26, 65, 75 e 80.

QUINLAN, R. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers, 1993. Citado na página 54.

RATTIGAN, M. J.; JENSEN, D. The case for anomalous link discovery. *SIGKDD Explor. Newsl.*, ACM, New York, NY, USA, v. 7, n. 2, p. 41–47, dez. 2005. ISSN 1931-0145. Disponível em: <http://doi.acm.org/10.1145/1117454.1117460>. Citado 2 vezes nas páginas 28 e 135.

RODRIGUEZ, J. J.; KUNCHEVA, L. I.; ALONSO, C. J. Rotation forest: A new classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 28, n. 10, p. 1619–1630, 2006. ISSN 0162-8828. Disponível em: <http://doi.ieeecomputersociety.org/10.1109/TPAMI.2006.211>. Citado na página 57.

RODRIGUEZ, M. G.; ROGATI, M. Bridging offline and online social graph dynamics. In: *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2012. (CIKM '12), p. 2447–2450. ISBN 978-1-4503-1156-4. Citado 3 vezes nas páginas 63, 75 e 80.

SA, H. de; PRUDENCIO, R. Supervised link prediction in weighted networks. In: *Neural Networks (IJCNN), The 2011 International Joint Conference on*. [S.l.: s.n.], 2011. p. 2281–2288. ISSN 2161-4393. Citado 5 vezes nas páginas 26, 36, 65, 74 e 134.

SALTON, G.; MCGILL, M. J. *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc., 1986. ISBN 0070544840. Citado na página 39.

SEEWALD, A. How to make stacking better and faster while also taking care of an unknown weakness. In: SAMMUT, C.; HOFFMANN, A. (Ed.). *Nineteenth International Conference on Machine Learning*. [S.l.]: Morgan Kaufmann Publishers, 2002. p. 554–561. Citado na página 55.

SHI, H. *Best-first decision tree learning*. Dissertação (Mestrado) — University of Waikato, Hamilton, NZ, 2007. COMP594. Citado na página 55.

SHIN, D.; SI, S.; DHILLON, I. S. Multi-scale link prediction. In: *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2012. (CIKM '12), p. 215–224. ISBN 978-1-4503-1156-4. Citado 2 vezes nas páginas 66 e 78.

SOARES, P. da S.; PRUDENCIO, R. B. C. Time series based link prediction. In: *Neural Networks (IJCNN), The 2012 International Joint Conference on*. [S.l.: s.n.], 2012. p. 1–7. ISSN 2161-4393. Citado 4 vezes nas páginas 36, 66, 72 e 134.

SONG, H. H. et al. Scalable proximity estimation and link prediction in online social networks. In: *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference*. New York, NY, USA: ACM, 2009. (IMC '09), p. 322–335. ISBN 978-1-60558-771-4. Citado 3 vezes nas páginas 20, 66 e 77.

SONG, H. H. et al. Clustered embedding of massive social networks. *SIGMETRICS Perform. Eval. Rev.*, ACM, New York, NY, USA, v. 40, n. 1, p. 331–342, jun. 2012. ISSN 0163-5999. Citado 2 vezes nas páginas 66 e 81.

SØRENSON, T. *A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons*. I kommission hos E. Munksgaard, 1948. Disponível em: <https://books.google.co.in/books?id=rpS8GAAACAAJ>. Citado na página 39.

STEURER, M.; TRATTNER, C. Predicting interactions in online social networks: An experiment in second life. In: *Proceedings of the 4th International Workshop on Modeling Social Media*. New York, NY, USA: ACM, 2013. (MSM '13), p. 5:1–5:8. ISBN 978-1-4503-2007-8. Citado 2 vezes nas páginas 66 e 78.

SU, J. et al. Discriminative parameter learning for bayesian networks. In: *Proceedings of the 25th International Conference on Machine Learning*. New York, NY, USA: ACM, 2008. (ICML '08), p. 1016–1023. ISBN 978-1-60558-205-4. Disponível em: <http://doi.acm.org/10.1145/1390156.1390284>. Citado na página 49.

SUMNER, M.; FRANK, E.; HALL, M. Speeding up logistic model tree induction. In: *9th European Conference on Principles and Practice of Knowledge Discovery in Databases*. [S.l.]: Springer, 2005. p. 675–683. Citado na página 50.

TIAN, Y. et al. Boosting social network connectivity with link revival. In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2010. (CIKM '10), p. 589–598. ISBN 978-1-4503-0099-5. Citado 2 vezes nas páginas 26 e 66.

TING, K. M.; WITTEN, I. H. Stacking bagged and dagged models. In: FISHER, D. H. (Ed.). *Fourteenth international Conference on Machine Learning*. San Francisco, CA: Morgan Kaufmann Publishers, 1997. p. 367–375. Citado na página 56.

TYLENDA, T.; ANGELOVA, R.; BEDATHUR, S. Towards time-aware link prediction in evolving social networks. In: *Proceedings of the 3rd Workshop on Social Network Mining and Analysis*. New York, NY, USA: ACM, 2009. (SNA-KDD '09), p. 9:1–9:10. ISBN 978-1-60558-676-2. Citado 2 vezes nas páginas 67 e 70.

VALVERDE-REBAZA, J.; LOPES, A. de A. Structural link prediction using community information on twitter. In: *Computational Aspects of Social Networks (CASoN), 2012 Fourth International Conference on*. [S.l.: s.n.], 2012. p. 132–137. Citado na página 67.

VASUKI, V. et al. Affiliation recommendation using auxiliary networks. In: *Proceedings of the Fourth ACM Conference on Recommender Systems*. New York, NY, USA: ACM, 2010. (RecSys '10), p. 103–110. ISBN 978-1-60558-906-0. Citado 2 vezes nas páginas 26 e 67.

VASUKI, V. et al. Scalable affiliation recommendation using auxiliary networks. *ACM Trans. Intell. Syst. Technol.*, ACM, New York, NY, USA, v. 3, n. 1, p. 3:1–3:20, out. 2011. ISSN 2157-6904. Disponível em: <http://doi.acm.org/10.1145/2036264.2036267>. Citado 2 vezes nas páginas 67 e 78.

WANG, C.; SATULURI, V.; PARTHASARATHY, S. Local probabilistic models for link prediction. In: *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*. [S.l.: s.n.], 2007. p. 322–331. ISSN 1550-4786. Citado na página 67.

WANG, D. et al. Human mobility, social ties, and link prediction. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2011. (KDD '11), p. 1100–1108. ISBN 978-1-4503-0813-7. Citado 3 vezes nas páginas 67, 76 e 77.

WANG, E. et al. Dynamic relational topic model for social network analysis with noisy links. In: *Statistical Signal Processing Workshop (SSP), 2011 IEEE*. [S.l.: s.n.], 2011. p. 497–500. ISSN pending. Citado na página 67.

WASSERMAN, S.; FAUST, K. *Social network analysis: Methods and applications*. [S.l.]: Cambridge university press, 1994. v. 8. Citado 3 vezes nas páginas 31, 32 e 34.

WEBB, G. Decision tree grafting from the all-tests-but-one partition. In: . San Francisco, CA: Morgan Kaufmann, 1999. Citado na página 54.

WEBB, G. I. Multiboosting: A technique for combining boosting and wagging. *Machine Learning*, Kluwer Academic Publishers, Boston, Vol.40, n. No.2, 2000. Citado na página 56.

WOLPERT, D. H. Stacked generalization. *Neural Networks*, Pergamon Press, v. 5, p. 241–259, 1992. Citado na página 55.

XIA, S. et al. Link prediction for bipartite social networks: The role of structural holes. In: *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*. [S.l.: s.n.], 2012. p. 153–157. Citado na página 67.

YIN, D.; HONG, L.; DAVISON, B. D. Structural link analysis and prediction in microblogs. In: *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2011. (CIKM '11), p. 1163–1168. ISBN 978-1-4503-0717-8. Citado na página 68.

YIN, Z. et al. A unified framework for link recommendation using random walks. In: *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on*. [S.l.: s.n.], 2010. p. 152–159. Citado na página 37.

YU, X. et al. Geo-friends recommendation in gps-based cyber-physical social network. In: *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*. [S.l.: s.n.], 2011. p. 361–368. Citado 2 vezes nas páginas 68 e 76.

ZAFARANI, R.; ABBASI, M.; LIU, H. *Social Media Mining: An Introduction*. [S.l.]: Cambridge University Press, 2014. ISBN 9781139916127. Citado 2 vezes nas páginas 33 e 34.

ZHANG, C.; ZHAI, B. Y.; WU, M. Link prediction of community in microblog based on exponential random graph model. In: *Wireless Personal Multimedia Communications (WPMC), 2013 16th International Symposium on*. [S.l.: s.n.], 2013. p. 1–6. ISSN 1347-6890. Citado na página 68.

ZHONG, E. et al. Modeling the dynamics of composite social networks. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2013. (KDD '13), p. 937–945. ISBN 978-1-4503-2174-7. Citado 2 vezes nas páginas 26 e 68.

ZHOU, T.; L, L.; ZHANG, Y.-C. Predicting missing links via local information. *The European Physical Journal B*, Springer-Verlag, v. 71, n. 4, p. 623–630, 2009. ISSN 1434-6028. Disponível em: <http://dx.doi.org/10.1140/epjb/e2009-00335-8>. Citado na página 39.