14 e 15 de outubro de 2021

Anais do 4º Workshop de Pesquisa em Computação dos Campos Gerais



WPCCG 2021

Gleifer V. Alves, André P. Borges, Erikson F. de Morais
Leandro M. Zatesko e Sheila M. de Almeida

Universidade Tecnológica Federal do Paraná - Ponta Grossa - PR

ISSN:2526-1371
DOI:10.5281/zenodo.6090322

Prefácio

Este volume contém os artigos apresentados no WPCCG 2021: 4^o Workshop de Pesquisa em Computação dos Campos Gerais, ocorrido nos dias 14 e 15 de outubro de 2021, na Universidade Tecnológica Federal do Paraná (UTFPR), em Ponta Grossa, Paraná, Brasil.

O WPCCG foi criado em 2016 para divulgar pesquisas em desenvolvimento ou concluídas de alunos e docentes de Instituições de Ensino Superior. O evento é organizado por docentes do Departamento Acadêmico de Informática (DAINF) dos Câmpus Ponta Grossa e Curitiba da UTFPR.

Nesta edição, pesquisadores foram convidados a submeter trabalhos relevantes para a área de Computação em duas categorias: resumos de uma página sobre pesquisa em andamento ou concluída; ou artigos com quatro a seis páginas apresentando resultados de pesquisa.

Em 2021, o Workshop teve 40 submissões. Cada submissão foi revisada por, ao menos, três revisores das áreas da Computação. Dos artigos submetidos, 35 foram aceitos para apresentação e publicação nesta edição. Historicamente, a participação da comunidade de Computação no WPCCG tem crescido e a taxa de aceitação das submissões é em média de 80%. Em números, a edição de 2016 teve 20 submissões (16 trabalhos aceitos), em 2017 foram 26 submissões (18 trabalhos aceitos), em 2019 foram 38 submissões (31 trabalhos aceitos) e em 2021 foram 40 submissões (35 trabalhos aceitos).

Outro fato importante a se destacar é a participação de autores de outras regiões do país, além do Paraná. Foram 76 autores de instituições paranaenses, mas também houve participações do Maranhão (2 autores), Mato Grosso do Sul (1 autor), Rio de Janeiro (14 autores), Santa Catarina (2 autores), São Paulo (3 autores), Alemanha(1 autor), Canadá (1 autor) e Portugal (1 autor).

Noventa revisores trabalharam nesta edição, dos quais 86 estão vinculados a instituições em estados brasileiros: Bahia (01), Goiás (01), Minas Gerais (03), Mato Grosso do Sul (06), Paraná (46), Rio de Janeiro (11), Rio Grande do Sul (04), Santa Catarina (04), São Paulo (10), e 4 estão em outros países: Argentina (01), Portugal (02) e Reino Unido (01). Ao todo, foram realizadas 122 revisões.

O gerenciamento de toda conferência foi feito por meio do sistema EasyChair.

O Comitê de Organização do WPCCG 2021 deixa aqui registrado um agradecimento especial aos 99 autores que escolheram este evento para divulgar suas pesquisas, e aos 90 revisores que colaboraram de maneira fundamental para a qualidade do evento.

04 de fevereiro de 2022
Ponta Grossa

Gleifer Vaz Alves
André Pinz Borges
Erikson Freitas de Morais
Leandro Miranda Zatesko
Sheila Morais de Almeida

Comitê de Programa

Alane Marie de Lima	Universidade Federal do Paraná - UFPR
Aleffer Rocha	Universidade Federal do Paraná - UFPR
Alessandra Hoffman	Universidade Tecnológica Federal do Paraná - UTFPR
André Koscianski	Universidade Tecnológica Federal do Paraná - UTFPR
André L. P. Guedes	Universidade Federal do Paraná - UFPR
André Pinz Borges	Universidade Tecnológica Federal do Paraná - UTFPR
Andre Roberto Ortoncelli	Universidade Tecnológica Federal do Paraná - UTFPR
Cândida Nunes da Silva	Universidade Federal de São Carlos - UFSCar
Carla Negri Lintzmayer	Universidade Federal do ABC - UFABC
Carlos Pantoja	Centro Federal de Educação Tecnológica do Rio de Janeiro - CEFET/RJ
Celina Figueiredo	Universidade Federal do Rio de Janeiro - UFRJ
Celso Aimbiré Weffort Santos	Universidade Estadual de Campinas - UNICAMP
Claudio Biazus	Universidade Comunitária da Região de Chapecó - UNOCHAPECÓ
Cleiton Santos	Universidade Federal do Paraná - UFPR
Daniel Costa de Paiva	Universidade Federal Fluminense - UFF
Daniel Posner	Universidade Federal do Rio de Janeiro - UFRJ
Dênis Lucas Silva	Universidade Tecnológica Federal do Paraná - UTFPR
Diana Francisca Adamatti	Universidade Federal do Rio Grande - FURG
Diana Sasaki	Universidade do Estado do Rio de Janeiro - UERJ
Edna Hoshino	Universidade Federal de Mato Grosso do Sul - UFMS
Emerson Paraiso	Pontifícia Universidade Católica do Paraná - PUCPR
Erikson Morais	Universidade Tecnológica Federal do Paraná - UTFPR
Evandro Kuszera	Universidade Tecnológica Federal do Paraná - UTFPR
Fabiano Oliveira	Universidade do Estado do Rio de Janeiro - UERJ
Fabio Protti	Universidade Federal Fluminense - UFF
Gabriel Sobral	Universidade de São Paulo - USP
Gedson Faria	Universidade Federal do Mato Grosso do Sul - UFMS
Gleifer Alves	Universidade Tecnológica Federal do Paraná - UTFPR
Guilherme Oliveira Mota	Universidade de São Paulo - USP
Helyane Borges	Universidade Tecnológica Federal do Paraná - UTFPR
Henrique Hepp	Universidade Federal do Paraná - UFPR
Hugo Kooki Kasuya Rosado	Universidade Estadual de Campinas - UNICAMP
Ionildo José Sanches	Universidade Tecnológica Federal do Paraná - UTFPR
Jean Paul Barddal	Pontifícia Universidade Católica do Paraná - PUC-PR
Jerusa Marchi	Universidade Federal de Santa Catarina - UFSC
Leandro Zatesko	Universidade Tecnológica Federal do Paraná - UTFPR
Lehilton L. C. Pedrosa	Universidade Estadual de Campinas - UNICAMP
Lívia Medeiros	Universidade do Estado do Rio de Janeiro - UERJ

Loana Tito Nogueira	Universidade Federal Fluminense - UFF
Lucas Castro	Universidade Estadual de Campinas - UNICAMP
Luciene Marin	Universidade Tecnológica Federal do Paraná - UTFPR
Luís Cassiano Goularte Rista	Pontifícia Universidade Católica do Rio Grande do Sul - PUC/RS
Luis Kowada	Universidade Federal do Rio de Janeiro - UFRJ
Luiz Carlos Pessoa Albini	Universidade Federal do Paraná - UFPR
Luiz Rafael Schmitke	Universidade Tecnológica Federal do Paraná - UTFPR
Maicon Rafael Zatelli	Universidade Federal de Santa Catarina - UFSC
Maiquel de Brito	Universidade Federal de Santa Catarina - UFSC
Marcelo Pinheiro Leite Benedito	Universidade Estadual de Campinas -UNICAMP
Márcia Cappelle Santana	Universidade Federal de Goiás - UFG
Marco Antonio Barbosa	Universidade Tecnológica Federal do Paraná - UTFPR
Marco Aurélio Graciotto Silva	Universidade Tecnológica Federal do Paraná - UTFPR
Marilton Aguiar	Universidade Federal de Pelotas - UFPEL
Marina Groshaus	Universidade Tecnológica Federal do Paraná - UTFPR
Marisa Gutierrez	Universidad Nacional de La Plata - UNLP
Murilo Da Silva	Universidade Federal do Paraná - UFPR
Nadia Kozievitch	Universidade Tecnológica Federal do Paraná - UTFPR
Neumar Malheiros	Universidade Federal de Lavras - UFLA
Priscila Martins	Universidade Federal de Mato Grosso do Sul - UFMS
Rafael C. Cardoso	The University of Manchester
Rafael Schouery	Universidade Estadual de Campinas - UNICAMP
Raphael Machado	INMETRO/Clavis
Renato Carmo	Universidade Federal do Paraná - UFPR
Richard Ribeiro	Universidade Tecnológica Federal do Paraná - UTFPR
Samuel Bellido Rodrigues	Universidade Tecnológica Federal do Paraná - UTFPR
Sheila Morais de Almeida	Universidade Tecnológica Federal do Paraná - UTFPR
Simone B. K. Aires	Universidade Tecnológica Federal do Paraná - UTFPR
Simone Matos	Universidade Tecnológica Federal do Paraná - UTFPR
Tarcizio Bini	Universidade Tecnológica Federal do Paraná - UTFPR
Uéverton Souza	Universidade Federal Fluminense - UFF
Vagner Pedrotti	Universidade Federal de Mato Grosso do Sul - UFMS
Vinícius Andrade	Universidade Tecnológica Federal do Paraná - UTFPR
Vinicius F. dos Santos	Universidade Federal de Minas Gerais - UFMG
Vitor T. F. de Luca	Universidade do Estado do Rio de Janeiro - UERJ

Sumário

Seção 1: Resumos

Geração de indicadores da Secretaria de Saúde de Curitiba com base no modelo de dados abertos conectados . . . 1
Wagner Luís Pereira e Nádia Puchalski Kozievitc . . .
Agrupamento de dados: um estudo de caso sobre registros de ocorrências em serviços de atendimento móvel às urgências . . . 2
Guilherme Vieira Bochi, Rosangela Villwock e Janaina Maria de Lima Gonçalves . . .
Geração de grafo de cena utilizando teorema de causa e efeito na extração e classificação das propriedades . . . 3
Flávio Henrique da Silva e André L. P. Guedes . . .
Simulação computacional na área de segurança pública . . . 4
Éliton José Laroca e André Koscianski . . .
Avaliação de arquiteturas de redes neurais siamesas para a classificação de espécies de plantas . 5
Matheus Moresco, Alceu de Souza Britto Jr. e Yandre Maldonado e Gomes da Costa . . .
Problemas de otimização utilizando o SCIP . . . 6
Viviane Frida Belli e André L. P. Guedes . . .
Fullerene nanodiscs: from Chemistry to Combinatorics . . . 7
Mariana Martins, Celina de Figueiredo, Diana Sasaki e Marcus Tovar . . .
Equitable total coloring of small graphs . . . 8
Matheus Adauto, Celina Figueiredo e Diana Sasaki . . .
Implementação de um jogo 3D com dados geográficos do OpenStreetMap . . . 9
Lucas do Nascimento Cadena e André Koscianski . . .
Evidências para a Conjectura De Hougardy . . . 10
Felipe Eduardo Pizzi, Juliano Silva do Nascimento e Sheila Morais de Almeida . . .
Simulation study based game . . . 11
André Koscianski, Guilherme Takeshi Saeki Abreu e Luiz Guilherme Monteiro Padilha . .
A realidade das abordagens adaptadas em problemas de otimização contínua . . . 12
Raphael Bezerra Xavier e André Luiz Pires Guedes . . .
Reconhecimento de grafos IIC-comparabilidade . . . 13
Lucas F. Glir e André Luiz Pires Guedes . . .
Integrando software e hardware em agentes inteligentes . . . 14
Vitor Luis Babireski Furio e Maiquel de Brito . . .
Sistema de previsão de preços e reserva de vagas em *smart parking* utilizando agentes . . . 15
Alexandre L. L. Mellado, Gleifer Vaz Alves, Paulo Leitão e André Pinz Borges . . .
Tendências para o desenvolvimento de jogos sérios na área de educação financeira . . . 16
Isabel Cristina Torrens, Simone Nasser Matos, Helyane Bronoski Borges e Rui Pedro Lopes
Um método para seleção de atributos em bases de dados de classificação hierárquica multirrótulo 17
Raimundo Osvaldo Vieira e Helyane Bronoski Borges . . .
Tessellation cover for bounded treewidth graphs . . . 18
Daniel Fabio Domingues Posner e Renan Guimarães da Silva . . .

Coloração de arestas em grafos indiferença 19
Rafael Porto e Sheila M. Almeida . . . . .
Minimal feedback arc sets containing long paths in few-vertex tournament 20
Fábio Botler e Rafael Schneider . . . . .
Flood-it game on co-comparability graphs 21
Guilherme Quintero Lorenzi e Sheila Morais de Almeida . . . . .
Coloração de arestas em grafos split proibidos para comparabilidade e intervalos 22
Cintia I. Cararo, Sheila M. Almeida, Cândida N. Silva e Glasielly D. Proença . . . . .
Deep learning aplicado à estimativa de profundidade em imagens estéreo multivista 23
Sérgio M. M. de Oliveira, Erikson F. de Morais e Marcella S. R. Martins . . . . .
Definição de pesos aos atributos em uma base de dados sem domínio definido no raciocínio baseado em casos 24
Jonathan William Hrenechen, André Koscianski e André Pinz Borges . . . . .
Cálculo de raízes quadradas de grafos utilizando um resolvedor pseudo-booleano 25
Pietro Cavassin, Aleffer Rocha e Renato Carmo . . . . .

Seção 2: Resumos estendidos

Configuração de parâmetro no algoritmo de agrupamento de dados baseado em formigas 26
Carina Moreira Costa e Rosangela Villwock . . . . .
A deductive-formal derivation for the preferential attachment metric for link prediction 31
Alane M. de Lima, Murilo V. G. da Silva e André L. Vignatti . . . . .
Criminalidade e pandemia: uma análise do impacto da COVID-19 na incidência de crimes em Curitiba 37
Patriky Edeovan Galvão Mirkoski . . . . .
An extended Adjacency Lemma for graph edge-coloring 43
Thiago H. F. M. Cunha e Leandro M. Zatesko . . . . .
Proper $\mathcal{L}(h,k)$-labelling of caterpillars and multisunlets 48
J. P. K. Castilho, C. N. Campos e L. M. Zatesko . . . . .
Grover's quantum algorithm and the maximum clique problem 54
Matheus Giovanni Dias e Leandro Miranda Zatesko . . . . .
On conformability of line graphs of complete graphs 60
Luerbio Faria, Mauro Nigro e Diana Sasaki . . . . .
Mapeamento sistemático sobre redes neurais convolucionais aplicadas na segmentação de imagens aéreas 65
Fabricio Bizotto e Mauren Louise Sguario Coelho De Andrade . . . . .
Desenvolvimento de um algoritmo de classificação de dados baseado na relevância da informação 71
Júlio César Guimarães Costa, Gleifer Vaz Alves e André Pinz Borges . . . . .
Imputação de dados faltantes em base de dados de classificação hierárquica multirrótulo usando regressão linear e polinomial 77
Alvaro Mateus Santana e Helyane Bronoski Borges . . . . .

Índice de Autores

Abreu, Guilherme Takeshi Saeki 11
Adauto, Matheus 8
Almeida, Sheila Morais de 10, 19, 21, 22
Alves, Gleifer Vaz 15, 71
Andrade, Mauren Louise Sguario Coelho De 65

Bizotto, Fabricio 65
Bochi, Guilherme Vieira 2
Borges, André Pinz 15, 24, 71
Borges, Helyane Bronoski 16, 17, 77
Botler, Fábio 20
Brito, Maiquel de 14
Britto Jr, Alceu de Souza 5

Cadena, Lucas do Nascimento 9
Campos, C. N. 48
Cararo, Cintia I. 22
Carmo, Renato 25
Castilho, J. P K. 48
Cavassin, Pietro 25
Costa, Carina Moreira 26
Costa, Júlio César Guimarães 71
Cunha, Thiago H. F. M. 43

Dias, Matheus Giovanni 54

Faria, Luerbio 60
Figueiredo, Celina de 7, 8
Furio, Vitor Luis Babireski 14

Glir, Lucas F. 13
Gomes da Costa, Yandre Maldonado 5
Gonçalves, Janaina Maria de Lima 2
Guedes, André L. P. 3, 12, 13

Hrenechen, Jonathan William 24

Koscianski, André 4, 9, 11, 24

Kozievitch, Nádia Puchalski 1

Laroca, Éliton José 4
Leitão, Paulo 15
Lima, Alane M. de 31
Lopes, Rui Pedro 16
Lorenzi, Guilherme Quintero 21

Martins, Marcella S. R. 23
Martins, Mariana 7
Matos, Simone Nasser 16
Mellado, Alexandre L. L. 15
Mirkoski, Patriky Edeovan Galvão 37
Morais, Erikson F. de 23
Moresco, Matheus 5

Nascimento, Juliano Silva do 10
Nigro, Mauro 60

Oliveira, Sérgio M. M. de 23

Padilha, Luiz Guilherme Monteiro 11
Pereira, Wagner Luís 1
Pizzi, Felipe Eduardo 10
Porto, Rafael 19
Posner, Daniel Fabio Domingues 18
Proença, Glasielly D. 22

Rocha, Aleffer 25

Santana, Alvaro Mateus 77
Sasaki, Diana 7, 8, 60
Schneider, Rafael 20
Silva, Cândida N. 22
Silva, Flávio Henrique 3
Silva, Murilo V. G. da 31
Silva, Renan Guimarães da 18

Torrens, Isabel Cristina 16
Tovar, Marcus 7

Vieira, Raimundo Osvaldo 17
Vignatti, André L. 31
Villwock, Rosangela 2, 26

Xavier, Raphael Bezerra 12

Zatesko, Leandro M. 43, 48, 54

Índice de Palavras-chave

Adaptações 12
Agentes 14
Agentes e sistemas multiagentes 15
Aprendizado profundo 23
Aprendizado profundo em grafos 3
Aprendizagem de máquina 15, 71
Automação veicular 23

Características multicamadas 5
Caterpillar graphs 48
Cidades inteligentes 1
Classificação de dados 71
Classificação de espécies de plantas 5
Classificação de imagens 3
Classificação de problemas de otimização 6
Classificação hierárquica multirrótulo 17, 77
Co-comparability graph 21
Coloração de arestas 19, 22
Coloração de vértices 10
Coloring of graphs and hypergraphs 43
Combinatorics 7
Complete graph 60
Computational complexity 18, 21
Conformable 60
Conjectura de Hougardy 10
Criminalidade 37
Cubic graphs 8
Curitiba 37

Dados abertos 1
Dados conectados 1
Dados faltantes 77
Data-driven 24
Decomposição Modular 13

Edge subsets with special properties 43
Educação financeira 16
Entity matching 1

Equitable total coloring 8
Estacionamentos inteligentes 15

Flood-it 21
Flooding game 21
Formal analysis 31
Fullerene 7

Game 11
Game engine 11
Geração de grafo de cena 3
Grafo 13
Grafo de comparabilidade 13, 22
Grafo de intervalos 19, 22
Grafo indiferença 19
Grafo split 22
Graph algorithms 43, 54
Graph coloring 18
Graph labelling 48
Graph theory 7, 18

Hibridizações 12

Imagens aéreas 65
Indicadores de saúde 1

K-means 2
KDD 2

Jogos de computador 9
Jogos sérios 16

Lichiardopol's conjecture 20
Line graph 60
Link prediction 31

Matriz de dissimilaridade 26
Metaheurística 26
Mineração de dados 2, 26, 71
Minimal feedback arc set 20

Nanodisc 7
Número cromático 10

OpenStreetMap 9
Ordem parcial 13
Orientação transitiva 13
Otimização 6, 12
OWL 1

Pandemia 37
Peso de atributos 24
Preferential attachment 31

Previsão de crimes 4
Programa pseudo-booleano 25

Quantum algorithms and complexity in the theory of computing 54
Quasi-paridade estrita 10

Raiz quadrada de grafos 25
RBC 24
RDF 1
Realidade aumentada 9
Redes neurais de grafos 3
Rede neural convolucional 65
Redes neurais siamesas 5
Redução de dimensionalidade 17
Regressão 37, 77
Regular graphs 8

Satisfazibilidade 25
SCIP 6
Segmentação Semântica 65
Segurança Pública 4
Seleção de atributos 17
Série temporal 37
Simulação 4
Simulation 11
Sistemas ciberfísicos 14
Sistemas embarcados 14
Sunlet graphs 48

Tendências 16
Total coloring 7, 60
Tournament 20

Vertex subsets with special properties 54
Visão estéreo multivista 23

GERAÇÃO DE INDICADORES DA SECRETARIA DE SAÚDE DE CURITIBA COM BASE NO MODELO DE DADOS ABERTOS CONECTADOS

Wagner Luís Pereira[a,b], Nádia Puchalski Kozievitch[a]

[a] *DAINF - Departamento Acadêmico de Informática - UTFPR, Av. Sete de Setembro, 3165, Curitiba, Paraná, Brasil*
[b] *Autor para correspondência: wagnerluis1989@gmail.com*

Palavras-chaves: Dados abertos, Dados Conectados, Indicadores de saúde, RDF, OWL, Cidades Inteligentes, *Entity Matching*

Relatórios de gestão são documentos elaborados para apresentar uma visão geral de desempenho, perspectivas, conectando e resumindo várias fontes de dados e indicadores. Apesar da Prefeitura de Curitiba possuir um portal de dados abertos, a secretaria municipal de saúde também gera informações que só estão disponíveis em relatórios quadrimestrais e anuais (como análises via indicadores). Os relatórios não possuem padronização (apesar de terem indicadores em comum), e geralmente são específicos a um público, dificultando a comparação de informações históricas e sendo dificilmente interpretáveis para outros fins. Esse fato traz à vista uma demanda de gerar um modelo para o armazenamento destes dados que traga maior visibilidade e comparabilidade. O presente projeto pretende mapear as fontes dos dados contidas nestes relatórios (como portais de informação de outros órgãos governamentais), mapeando indicadores que se mantêm ao longo do tempo, utilizando conceitos nas áreas de Banco de dados, Cidade Inteligentes e Inteligência Artificial. O objetivo é algoritmos de *Entity Matching*, que são capazes de localizar termos semelhantes, para localizar indicadores que se mantêm ao longo do tempo, porém tem variações em seus títulos. Além disso será proposto um modelo de dados no padrão de Dados Abertos Conectados, com uma parte dos indicadores mapeados, com o objetivo de apresentar a viabilidade da construção de um modelo que abranja todos os dados e que possa ser utilizado por qualquer pessoa ou aplicação.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Agrupamento de Dados: um estudo de caso sobre registros de ocorrências em Serviços de Atendimento Móvel às Urgências

Guilherme Vieira Bochi[a,b], Rosangela Villwock[a], Janaina Maria de Lima Gonçalves[a]

[a] *Universidade Estadual do Oeste do Paraná - UNIOESTE*
[b] *Autor para correspondência: guilhermebochi@gmail.com*

Palavras-chaves: KDD, Mineração de Dados, *K-means*

O processo KDD (Knowledge Discovery in Database), que significa Descoberta de Conhecimento em Base de Dados, é utilizado na busca de um conhecimento novo, útil ou válido em base de dados. Este trabalho teve como objetivo explorar uma base de dados do SAMU (Serviços de Atendimento Móvel às Urgências) por meio do processo KDD, buscando avaliar o perfil de cada grupo formado, bem como relacionar as características dos grupos com os tipos de agravo. Para isso, foi utilizada a tarefa de agrupamento de dados, com a utilização do algoritmo *K-means*. Na base de dados utilizada foram observados atributos de 19481 ocorrências atendidas no período de Março a Agosto do ano de 2020, na macro região do oeste do Paraná. Alguns dos atributos observados foram: o dia do mês em que a ocorrência se deu; o dia da semana em que houve a ocorrência; o horário em que o recurso foi acionado pra prestar o atendimento; a unidade móvel enviada até o local da ocorrência; a idade do paciente que foi assistido; o sexo do paciente assistido e o encaminhamento dado ao paciente. Para a implementação foi utilizada a função KMeans da biblioteca cluster do pacote scikit-learn, do Python. Observando os centróides dos sete grupos formados, observou-se que a idade dos indivíduos mostrou-se um fator relevante para a divisão dos grupos, enquanto o dia da semana não. Em alguns grupos as ocorrências se concentram mais no início do mês e em outros grupos mais ao final do mês. Em alguns grupos houve maior ou menor incidência de ocorrências em determinados horários. A unidade móvel mais utilizada em todos os grupos foi a de suporte básico. Porém, para alguns grupos houve maior ou menor incidência de uso de unidades de suporte avançado. Quanto ao sexo, alguns grupos tem maior incidência de atendimentos do sexo masculino e outros do sexo feminino, porém, nenhum dos grupos com maioria expressiva de algum dos gêneros. Em quase todos os grupos a maior parte dos encaminhamentos foi hospitalar. Porém, para alguns grupos houve maior ou menor incidência de óbitos, por exemplo. Foi observada ainda a relação entre cada tipo de agravo com os grupos formados. Para o agravo **acidente de trânsito** maiores incidências foram observadas em grupos onde a maior parte dos atendimentos foi de adultos jovens e para o agravo **psiquiátrico** maiores incidências foram observadas em grupos que possuem atendimentos, em sua maioria, dos 11 aos 40 anos. Espera-se que os resultados obtidos possam servir de auxílio para tomadas de decisão no SAMU.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Geração de grafo de cena utilizando teorema de causa e efeito na extração e classificação das propriedades

Flávio Henrique da Silva[a], André L. P. Guedes[a]

[a]*Departamento de Informática, Universidade Federal do Paraná, Centro Politécnico, Curitiba, PR, Brasil*
[b]*fhsilva@inf.ufpr.br, andre@inf.ufpr.br*

Palavras-chaves: Geração de Grafo de Cena, Redes Neurais de Grafos, Aprendizado Profundo em Grafos, Classificação de Imagens

A utilização de grafos está presente em vários domínios do conhecimento e em diversas aplicações, desde a análise social, na fabricação de novos medicamentos até mesmo em Visão Computacional. Nesta última área as abordagens evidenciadas nos últimos anos na literatura buscam a descrição e a representação semântica de uma cena visual (imagem) independente da complexidade, níveis semânticos, tipos, atributos e quantidade de elementos que compõem a cena. Esta pesquisa tem como objetivo propor uma solução para as limitações descritas no processo de *Scene Graph Generation* (SGG) que compreendem desde da detecção dos relacionamentos, atributos, na construção de uma associação coerente entre os nós e vértices de um grafo heterogêneo de uma cena visual complexa e no tratamento do viés em grandes conjuntos de dados. Limitações que se encontram especificamente na etapa de refinamento das características dos nós que formam um grafo, o qual, por sua vez, representa uma imagem de diferentes níveis semânticos. A abordagem proposta visa combinar os métodos *Causal-TDE* (Tang et al. 2020) e *GPS-Net* (Xin et al. 2020), ambos apresentados na *Conference on Computer Vision and Pattern Recognition* (CVPR). O primeiro consiste na aplicação da teoria de causa e efeito no tratamento de viés em grandes conjuntos de dados durante a etapa de treinamento de um modelo através de técnicas de aprendizado profundo. Abordagem que será combinada com o processo de extração das características mais relevantes de um nó, com base na identificação e classificação das propriedades particulares de cada vértice. Por fim, o modelo construído será validado em conjuntos de imagens públicos, tais como: *Visual Genome*, *OpenImages* e *Visual Relationship Detection*, e os resultados serão comparados e analisados com demais métodos disponíveis na literatura, comparando assim a eficiência, precisão e robustez do modelo de SGG por meio da métrica de *Recall* (Kynkäänniemi et al. 2019).

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Simulação Computacional na Área de Segurança Pública

Éliton José Laroca[a], André Koscianski[a]

[a] *Departamento Acadêmico de Informática, Universidade Tecnológica Federal do Paraná, Rua Dr. Washington Subtil Chueire, 330, Ponta Grossa, Paraná, Brasil*

eliton@alunos.utfpr.edu.br, koscianski@utfpr.edu.br

Palavras-chaves: Simulação, Previsão de Crimes, Segurança Pública

Entre as inúmeras formas como os computadores podem auxiliar a sociedade, pode se destacar contribuições na área de segurança pública. Bancos de dados geográficos, por exemplo, permitem registrar dados espaço-temporais sobre crimes, a partir dos quais é possível realizar análises matemáticas e estatísticas. Ferramentas como análise de séries temporais abrem a possibilidade de otimizar a distribuição de contingentes policiais e reduzir custos. Além disso, a informatização pode contribuir para uma maior interação entre as organizações policias e de segurança do Estado.

Modelos de simulação computacional tem o potencial de auxiliar a compreender melhor a dinâmica de uma região, em função de variáveis como atratividade de um local, circulação de pessoas, iluminação, etc. A partir disso pode-se projetar cenários de administração, como escalas e rotas de patrulhamento, e investimentos em segurança.

Para construir uma simulação computacional, é preciso ter um modelo inicial e, no presente caso, um conjunto de dados que retrate crimes reais. Algumas características exploradas na literatura incluem explicar porque muitas vezes os crimes estão fortemente concentrados em pontos críticos, ou porque as vítimas se repetem, e como os crimes podem aumentar ou diminuir em determinada hora do dia, dia da semana ou até estação do ano. Algumas das técnicas usadas incluem regressão logística, redes neurais, e a simulação baseada em agentes computacionais.

O projeto está em estágio inicial e conta com a parceria da Polícia Militar do Paraná, para obtenção de dados reais de ocorrências que aconteceram no estado. Tais dados deverão ser integrados em um modelo de simulação baseado em agentes, sendo essa integração entre comportamentos e dados reais, um dos principais problemas de investigação.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

AVALIAÇÃO DE ARQUITETURAS DE REDES NEURAIS SIAMESAS PARA A CLASSIFICAÇÃO DE ESPÉCIES DE PLANTAS

Matheus Moresco[a,1], Alceu de Souza Britto Jr.[a], Yandre Maldonado e Gomes da Costa[b]

[a]*Mestrado em Computação Aplicada, Universidade Estadual de Ponta Grossa (UEPG), Uvaranas, Ponta Grossa, Paraná, Basil*
[b]*Departamento de Informática, Universidade Estadual de Maringá (UEM), Av. Colombo, 5790, Maringá, Paraná, Brasil*
[c]*Autor para correspondência: morescomatheus77@gmail.com*

Palavras-chaves: Redes Neurais Siamesas, Classificação de espécies de plantas, Características multicamadas

A classificação de plantas, em especial a que utiliza a imagem da folha, é tida como um grande desafio, devido ao grande número de espécies. Com o intuito de facilitar o processo, sistemas automáticos que usam técnicas de aprendizagem de máquina e visão computacional para diferenciar as espécies têm sido propostos, em especial modelos de Redes Neurais Convolucionais (CNN, do inglês, *"Convolutional Neural Networks"*), vem sendo amplamente utilizados para extração de características e classificação de espécies de plantas. Porém, o uso de tal ferramenta não permite a escalabilidade do modelo, de maneira que, caso uma nova classe seja adicionada, a rede precisa ser treinada novamente. Assim, uma alternativa interessante tem sido o uso de Redes Neurais Siamesas (SNN, do inglês, "*Siamese Neural Network*"), que diferente de uma CNN fornece uma medida de similaridade entre pares de imagens, permitindo um processo de classificação a partir de imagens de referência. Dessa forma, a inclusão de uma nova espécie de planta no processo demanda apenas a adição de ao menos uma imagem de referência. A partir disso, este estudo tem por objetivo avaliar diferentes arquiteturas de SNNs na classificação de espécies de plantas a partir da imagem da folha. Adicionalmente, pretende-se investigar a combinação de características extraídas de camadas intermediárias das SNNs, observando o impacto na classificação e buscando priorizar abordagens que sejam eficientes mesmo com poucas imagens de treinamento. Experimentos realizados com as bases de imagens de folhas *Flavia* e considerando na estrutura da SNN as arquiteturas *DenseNet* e *VGG-16*, demonstram que a combinação de características extraídas de camadas intermediárias é promissora. O melhor resultado (99,82%) foi observado com o uso da *DenseNet*, e representou uma melhora de 0,35 pontos percentuais quando comparada a SSN convencional (99,47%) e 1,22 quando comparada a CNN-*DenseNet* (98,6%).

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Problemas de Otimização Utilizando o SCIP

Viviane Frida Belli[a,b], André L. P. Guedes[a]

[a] *Departamento de Informática, UFPR, Curitiba, PR, Brasil*
[b] *Autor para correspondência: vfbelli@inf.ufpr.br*

Palavras-chaves: otimização, classificação de problemas de otimização, SCIP

A utilização de softwares para a resolução de problemas de otimização é prática recorrente nas mais diversas aplicações, desde simples problemas de determinação de rotas em projetos de mobilidade de cidades, até problemas de design de chips para a indústria de semicondutores. Considerando as diferentes categorias de problemas de otimização, existem métodos e técnicas específicos para sua resolução. Portanto, classificar o tipo de problema implica na possibilidade de utilização da ferramenta adequada para sua resolução. Dentre as classificações possíveis, por exemplo, podemos considerar a relação de linearidade ou não-linearidade das restrições e função objetivo. E nesta classificação existem categorias de problemas de otimização que possuem características em comum, mas também com atributos que as diferenciam. Estas categorias podem ser representadas em uma hierarquia na qual os métodos de resolução aplicados podem ser os mesmos ou totalmente diversos entre si. Por exemplo, um problema de programação convexa é também um problema de programação linear, em que especificamente a função objetivo e as restrições são convexas. A análise da eficiência dos métodos utilizados para a resolução de determinado tipo de problema permite demonstrar a aplicabilidade da ferramenta escolhida para o tipo de problema abordado. O SCIP (*Solving Constraint Integer Programs*) é um resolvedor não comercial para problemas de programação inteira mista (MIP: *Mixed Integer Programming*) e problemas de programação não-linear inteira mista (MINLP: *Mixed Integer Nonlinear Programming*), além de ser um framework para resolução de problemas de programação por restrições inteira (CIP: *Constraint Integer Program*). Desenvolvido pelos membros do Zuse Institute Berlin (ZIB) e atualmente em sua versão 7.0, o SCIP permite que o usuário tenha total controle e acesso a informações detalhadas do processo de resolução dos problemas. O objetivo deste trabalho é organizar uma possível classificação dos tipos de problemas de otimização, de acordo com critérios específicos, possibilitando que, dado um problema, este seja corretamente classificado. Feita a classificação, busca-se analisar quais métodos o SCIP utiliza na resolução dos diferentes tipos de problemas e para quais destes tipos sua utilização é eficiente.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Fullerene nanodiscs: from Chemistry to Combinatorics

Mariana Martins[a,c], Celina de Figueiredo[a], Diana Sasaki[b], Marcus Tovar[b]

[a] *PESC, COPPE, Rio de Janeiro Federal University, Cidade Universitária, Rio de Janeiro RJ, Brazil*
[b] *IME, Rio de Janeiro State University, Maracanã, Rio de Janeiro RJ, Brazil*
[c] *Corresponding author: mmartins@cos.ufrj.br*

Keywords: Graph Theory, Combinatorics, Total Coloring, Fullerene, Nanodiscs

A graph is a mathematical model used to represent relationships between objects. The general characteristics that objects and their relationships can assume allowed the construction of the (so-called) Graph Theory, which has been applied to model large scale problems in several areas, such as Mathematics, Physics, Computer Science, Engineering, Chemistry and Psychology.

Fullerene graphs are mathematical models for carbon-based molecules experimentally discovered in the early 1980. Many parameters associated with these graphs have been discussed to describe the stability of fullerene molecules. By definition, fullerene graphs are cubic, planar, 3-connected with pentagonal and hexagonal faces. A total coloring of a graph assigns colors to the vertices and edges of a graph such that adjacent or incident elements have different colors. The famous Total Coloring Conjecture opened at 50 years, is settled for cubic graphs, but not to arbitrary regular graphs nor to arbitrary planar graphs. The length of the shortest cycle in a graph is called girth. Our goal is to study the total coloring of an infinite subfamily of fullerene graphs, the fullerene nanodiscs D_r, with distance between the inner (outer) layer and the central layer given by the radius parameter $r \geq 2$, motivated by a conjecture that the girth of a graph is a relevant parameter in the study of total coloring. To highlight the choice of the studied graph class, we present a historical scenario of the carbon molecule discovery that can be modeled through a special cubic planar graph of girth 5.

We provide a complete combinatorial description of the small fullerene nanodiscs. We prove that there is just one fullerene nanodisc with radius 2, just one fullerene nanodisc with radius 3, and just two fullerene nanodiscs with radius 4. We contribute by giving the first combinatorial description for fullerene nanodiscs, aiming to improve the understanding of this class, so that we are able to tackle the challenging total coloring of these graphs.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Equitable Total Coloring of Small Graphs

Matheus Adauto[a,c], Celina Figueiredo[a,d], Diana Sasaki[b,1]

[a] *Federal University of Rio de Janeiro, Rio de Janeiro, Brazil*
[b] *Rio de Janeiro State University, Rio de Janeiro, Brazil*
[c] *adauto@cos.ufrj.br*
[d] *celina@cos.ufrj.br*
[e] *diana.sasaki@ime.uerj.br*

Keywords: regular graphs, cubic graphs, equitable total coloring

A *total k-coloring* of a simple graph $G = (V, E)$ assigns at most k colors to the vertices and edges of G such that distinct colors are assigned to every pair of adjacent vertices in V, to every pair of adjacent edges in E, and each vertex and its incident edges. A total coloring is *equitable* if the cardinalities of any two color classes differ by at most 1. The *total chromatic number* $\chi''(G)$ is the minimum k such that G admits a total k-coloring. Behzad and Vizing independently conjectured that for every simple graph, $\Delta + 1 \leq \chi''(G) \leq \Delta + 2$, know as the *Total Coloring Conjecture*, where Δ is the maximum degree in G. Graphs that satisfy $\chi''(G) = \Delta + 1$ are called *Type 1*.

In 2020, Stemock considered equitable total colorings of cubic graphs in his article "On the equitable total $(k + 1)$-coloring of k-regular graphs". The author conjectured that every total 4-coloring of a cubic graph is equitable if $n < 20$. This upper bound was motivated by a graph with $n = 20$ in the article "On the equitable total chromatic number of cubic graphs", by Dantas, Figueiredo, Mazzuoccolo, Preissmann, dos Santos and Sasaki, that is Type 1 but does not have an equitable total coloring with 4 colors. This conjecture becomes relevant when we realize that it refers to more than 40000 graphs, which can be verified in the website "House of Graphs". We found counter examples for the conjecture with $n = 12$ and 18 in the article "Some refinements of the total chromatic number conjecture", by Chetwynd and Hilton, where non equitable total colorings of these Type 1 graphs are described.

So far all small Type 1 cubic graphs we have analyzed have at least one equitable total 4-coloring. We propose instead the following conjecture: every Type 1 cubic graph has at least one equitable total 4-coloring for $n < 20$.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Implementação de um Jogo 3D com dados geográficos do OpenStreetMap

Lucas do Nascimento Cadena[a,b], André Koscianski[a]

[a]*Universidade Tecnológica Federal do Paraná, Rua Doutor Washington Subtil Chueire, 330, Ponta Grossa, Paraná, Brasil*
[b]*Autor para correspondência: lucascdna@gmail.com*

Palavras-chaves: Jogos de Computador, OpenStreetMap, Realidade Aumentada

Jogos de computador requerem um bom nível de programação e criatividade. São complicados pois é necessário se preocupar com diversos fatores que compõe um jogo. Por esse motivo necessitam de bibliotecas para seu desenvolvimento que são responsáveis por esses fatores como física, som, renderização, input do usuário, etc.

Um jogo muito famoso e de destaque foi Pokémon GO. Nele os jogadores são incluídos em um ambiente que mescla mundo simulado e real, através das funcionalidades de capturar criaturas utilizando realidade virtual, e a renderização de seu mapa que é feita através de dados geográficos reais. Uma forma de obter essa informação é pela plataforma OpenStreetMap que é disponibilizada de forma pública e colaborativa.

Neste projeto, o objetivo é importar o mapa de uma cidade que será cenário para um jogo que contará com objetivos, renderizações em 3D e inimigos controlados pelo computador. A ideia geral do jogo é simular um ambiente de sobrevivência em mundo aberto onde o principal objetivo é se manter vivo enquanto completa as missões disponíveis. Inicialmente o software será desenvolvido para computador e posteriormente avaliada uma versão Android. O projeto será uma plataforma de base para, futuramente, explorar opções como jogos envolvendo realidade virtual.

Dentro do trabalho foi implementada a importação das ruas, com base em um projeto existente de código aberto, através do Web Service do OpenStreetMap. A partir dessa implementação, foram necessários testes para encontrar a escala do mundo real dentro do projeto e o posicionamento de objetos a partir de latitude e longitude dentro do plano renderizado.

Para os próximos passos deste projeto, haverá a implementação da lógica geral do jogo, ou seja, como serão os objetivos e quais serão os inimigos, seus comportamentos e níveis de dificuldade.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Evidências para a Conjectura De Hougardy

Felipe Eduardo Pizzi[a,b], Juliano Silva do Nascimento[a], Sheila Morais de Almeida[a]

[a] *Departamento Acadêmico de Informática, Universidade Tecnológica Federal do Paraná, R. Doutor Washington Subtil Chueire, 330, Ponta Grossa – PR, Brasil*
[b] *Autor para correspondência: felipeeduardo@alunos.utfpr.edu.br*

Palavras-chaves: coloração de vértices, número cromático, conjectura de Hougardy, quasi-paridade estrita

O número cromático de um grafo G, $\chi(G)$, é o menor número de cores com que se pode colorir os vértices de G de maneira que vértices vizinhos tenham cores diferentes. Dado um grafo G e um número natural k, decidir se $\chi(G) \leq k$ é um problema NP-completo (E. Karp: *Reducibility among combinatorial problems*).

Uma clique é um conjunto de vértices dois a dois adjacentes. O tamanho da maior clique de um ygrafo G é denotado por $\omega(G)$. Um grafo G é perfeito se e somente se $\omega(H) = \chi(H)$ para todo subgrafo induzido H do grafo G. Uma dupla-par é um par de vértices não adjacentes tais que todos os caminhos induzidos entre eles têm comprimento par. A contração de duplas-pares em um grafo perfeito G produz um grafo perfeito H com o mesmo tamanho de maior clique. Portanto, $\chi(H) = \chi(G)$.

Um grafo é quasi-paridade estrita (SQP) se todo subgrafo induzido possui uma dupla-par ou é uma clique. Muitas subclasses de grafos perfeitos são quase-paridade estrita, como cografos, fracamente cordais, Gallai, Meyniel, e perfeitamente contraíveis. Mas, há grafos perfeitos que não têm dupla-par, como os grafos linha de bipartidos 3-conexos. Um grafo que não é SQP cujos subgrafos induzidos próprios são SQP é chamado de não-SQP minimal. Hougardy conjecturou que um grafo é não-SQP minimal se e somente se é um ciclo ímpar com pelo menos 5 vértices ou o complemento de um ciclo ímpar com pelo menos 7 vértices, ou um grafo linha de bipartido (S. Hougardy, *Even and odd pairs in linegraphs of bipartite graphs.*).

A Conjectura de Hougardy é verdadeira para os grafos livres de garras (C. L, Sales e F. Maffray, *Even Pairs in Claw-Free Perfect Graphs*); livres de diamantes (A. E. Kézdy e M. Scobee: *A proof of Hougardy's conjecture for diamond-free graphs*); livres de touro (C. H. M . de Figueiredo, F. Mafray e C. R. V. Maciel, *Even Pairs in Bull-reducible Graphs*); e planares (C. L. Sales, F. Maffray and B. A. Reed, *On planar quasi-parity graphs*).

Neste projeto, procuramos novas evidências para a Conjectura de Hougardy em uma superclasse dos grafos livres de touro e em uma superclasse dos grafos livres de diamante. A primeira é a classe dos grafos que não contém S_3, S_4 e $\overline{S_3}$ como subgrafos induzidos. A segunda é classe dos grafos probe livres de diamante.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Simulation Study Based Game

André Koscianski[a,b], :Guilherme Takeshi Saeki Abreu[c], Luiz Guilherme Monteiro Padilha[d]

[a]*DAINF, R.Doutor Washington Subtil Chueire, 330, Ponta Grossa, PR, Brazil*
[b] *Corresponding Author: koscianski@utfpr.edu.br*
[c] *Corresponding Author: TakeshiSaeki@outlook.com*
[C]*orresponding Author: corvonauta.dev@gmail.com*

Keywords: simulation, game, game engine

Many computer games are based on an underlying simulation. Some examples are Sid Meier' s Civilization, Europa Universalis and F1. In all these examples, the accuracy of the simulation, either regarding historical facts or physical equations, tuned according to the game goals.

In this study, that reasoning is inverted, giving space to a developing idea of a game with a basis in a simulation instead. The project will begin by developing a simulation in public security, with characteristics and behavior of the simulated environment will be the starting point for designing a computer game. The project will emphasize the design and development of software: the simulation model, game mechanics, and algorithms, and accord a little less priority to aesthetic factors. This approach considers that the focus of this study aims at a functional prototype at least, instead of a fully developed game.

The simulator to be developed will be a simplified model for the occurrence of crimes in a city. This theme is being studied by a Master' s degree project belonging to PPGCC, aiming to use real data. In the present project, a simplified model will be developed, considering requisites and indications from the master' s thesis. The execution of the simulator is expected to reflect, to some extent, actual scenarios, and data. The game mechanics and story will be designed considering the output of the simulation.

In the current phase of project development, preliminary studies were completed on the use of game engines, the design and implementation of simulators and games, and the game creation process. Along with the preliminary studies, the simulator modeling began. Given the data used for the simulator, revolving a simulated urban setting, encompassing both people and agents common to public environments.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

A Realidade das Abordagens Adaptadas em Problemas de Otimização Contínua

Raphael Bezerra Xavier[a,b], André Luiz Pires Guedes[a]

[a]*Departamento de Informática, UFPR, Curitiba, PR, Brasil*
[b]*Autor para correspondência: rbxavier@inf.ufpr.br*

Palavras-chaves: Otimização, Adaptações, Hibridizações

Comumente pesquisadores têm a necessidade de resolver problemas de otimização de alto grau de complexidade, como o de escalonamento de tarefas, alocação de pessoas e transporte. A solução destes problemas requer o desenvolvimento de algoritmos que produzam soluções com qualidade e que sejam obtidas com custo computacional reduzido.

O estudo de métodos híbridos tem se fortificado no intuito de minimizar o esforço computacional, porém existe uma tendência muito grande de autores continuarem crescendo o número de novas implementações destes métodos híbridos. Os algoritmos bio-inspirados estão na lista dos que mais sofrem adaptações, por exemplo, na aplicação em um problema de otimização de natureza contínua, (Y. C. Liang, J. R. C. Juarez: *A self-adaptive virus optimization algorithm for continuous optimization problems*) é proposto um algoritmo híbrido com um mecanismo autoadaptativo para diminuir o número de parâmetros controláveis. Com isto sustenta-se a ideia do estudo analítico para desmistificar e explicar o motivo de uma alta quantidade de adaptações e hibridizações em algoritmos de otimização.

O avanço deste estudo pode trazer razões e novas tendências para se analisar as melhores formas de como utilizar os algoritmos ou técnicas de otimização. Será realizado um levantamento bibliográfico, onde os problemas em que estes algoritmos são utilizados serão vistos de uma forma em que técnicas mais fundamentais podem tratá-los de forma mais simples.

Em (B. Nogueira, E. Tavares, J. Araujo, G. Callou: *Accelerating continuous GRASP with a GPU.*), por exemplo, usa-se uma versão do GRASP contínuo(C-GRASP) com uma paralelização utilizando *Graphics Processing Unit*(GPU) em suas duas fases, a fim de conseguir a redução do número de avaliações na execução do algoritmo.

Com isso conclui-se que é mais comum realizar uma nova adaptação para resolver um problema, do que tentar ver como o algoritmo em sua forma original pode se comportar e também resolver o mesmo problema. Será que este número alto de adaptações traz algum benefício computacional, ou os problemas simples são tratados de forma complexa para poder acelerar a convergência através destas adaptações?

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Reconhecimento de grafos IIC-comparabilidade

Lucas F.Glir[a,b], André L. P. Guedes[a]

[a] *Departamento de Informática, UFPR, Curitiba, PR, Brasil*
[b] *Autor para correspondência: lfglir@inf.ufpr.br*

Palavras-chaves: Grafo, Grafo de Comparabilidade, Ordem Parcial, Decomposição Modular, Orientação Transitiva

Dado um conjunto parcialmente ordenado (poset), $P = (C, \leq)$, o grafo de comparabilidade de P é o grafo $G_p = (V, E)$ tal que $uv \in E$ se e somente se $u \leq v$ ou $v \leq u$. Um grafo G é um grafo de comparabilidade se existe P tal que G e G_p são isomorfos. Dado um poset $P = (C, \leq)$ e $x \in C$, $I_p^-(x) = \{y \in C | y \leq x\}$ e $I_p^+(x) = \{y \in C | x \leq y\}$ são, respectivamente, os intervalos de predecessores e de sucessores de x em P. Dizemos que P é fechado por interseção em intervalos (IIC) se $\mathcal{I}_p^- = \{I_p^-(x) | x \in C\}$ e $\mathcal{I}_p^+ = \{I_p^+(x) | x \in C\}$ são fechados por interseção. Um grafo G é IIC-comparabilidade se existe um poset IIC P tal que G e G_p são isomorfos. Queremos estudar o problema de reconhecimento de grafos IIC-comparabilidade. Um grafo de comparabilidade sem C_4 induzido é IIC. Mas, é necessário que todo C_4 esteja em uma 4-roda (um C_4 mais um vértice universal). Isso não é condição suficiente, pois ainda é preciso encontrar uma orientação de forma que o vértice central fique no meio. Note que um poset sobre o conjunto de vértices é equivalente a uma orientação transitiva das arestas. Uma forma de atacar o problema de reconhecimento de grafos IIC-comparabilidade é através dos algoritmos de reconhecimento de grafos de comparabilidade. Golumbic, em 1977, propõe que se oriente cada aresta individualmente, garantindo que a orientação continue coesa durante a execução do algoritmo. Entretanto, é difícil saber por qual aresta deve-se começar para que o algoritmo oriente corretamente as 4-rodas. Outra maneira é através da orientação transitiva dos grafos quocientes resultantes de uma decomposição modular. Para orientar as arestas, se orienta o grafo quociente e em seguida se repete o processo para os grafos induzidos pelos módulos. Com isso se espera que seja possível testar se o grafo é ou não IIC-comparabilidade através das possibilidades de orientação.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Integrando software e hardware em agentes inteligentes

Vitor Luis Babireski Furio[a,b], Maiquel de Brito[a]

[a]*Departamento de Engenharia de Controle, Automação e Computação, Universidade Federal de Santa Catarina, Rua João Pessoa, 2750, Blumenau, SC, Brasil*
[b]*Bolsista PIBIC – Propesq/UFSC/CNPq – vitorluis.babireskifurio@gmail.com*

Palavras-chaves: agentes, sistemas embarcados, sistemas ciberfísicos

Agentes são entidades computacionais que apresentam autonomia, pró-atividade, reatividade e sociabilidade. Estas características são relevantes em sistemas ciberfísicos, que integram *hardware* e *software* para atuar sobre o ambiente físico em cenários complexos típicos em áreas como Indústria 4.0 e cidades inteligentes. Entretanto, a programação orientada a agentes (AOP) trata agentes como elementos exclusivamente de *software*. Sensores e atuadores são vistos como elementos externos aos agentes. A transformação de valores lidos pelos sensores em crenças deve ser programada no código do agente e as ações habilitadas pelos atuadores são tratadas como ações em um elemento externo. A implementação do agente mistura aspectos cognitivos com a manipulação do *hardware*. Assim, propõe-se uma extensão para o interpretador da linguagem de AOP *Jason* que permita conceber um agente como uma união de *software* e *hardware*. A extensão é concebida em camadas de *hardware* e *software* citadas a seguir de baixo para cima. As camadas de *hardware* são: componentes eletrônicos (sensores e atuadores), microcontrolador (controle de sensores e atuadores) e processador central (execução do código do agente). As camadas de *software* são: código do microcontrolador (controle dos componentes eletrônicos), comunicação física (comunicação entre microcontrolador e processador central), interface microcontrolador (representação do microcontrolador dentro do processador central), interface do agente (interface entre as camadas inferiores e código do agente) e código do agente. Esta arquitetura integra os níveis cognitivo e físico em um agente embarcado em um sistema ciberfísico que pode, inclusive, ser formado por vários microcontroladores. Por exemplo, um carro autônomo teria um processador central executando o código do agente e vários microcontroladores conectados a diferentes partes do veículo (ar condicionado, sistema de direção). Um único agente controlaria todo o veículo, podendo interagir com outros veículos (agentes).

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Sistema de Previsão de Preços e Reserva de Vagas em Smart Parking utilizando Agentes

Alexandre L. L. Mellado[a], Gleifer Vaz Alves[a], Paulo Leitão[b], André Pinz Borges[a]

mellado@alunos.utfpr.edu.br, gleifer@utfpr.edu.br, pleitao@ipb.pt, apborges@utfpr.edu.br

[a]*Departamento Acadêmico de Informática (DAINF), Universidade Tecnológica Federal do Paraná (UTFPR), Ponta Grossa, PR, Brasil*
[b]*Research Centre in Digitalization and Intelligent Robotics (CeDRI), Instituto Politécnico de Bragança (IPB), Bragança, Portugal*

Palavras-chaves: Agentes e Sistemas Multiagentes, Estacionamentos Inteligentes, Aprendizagem de Máquina

O problema de mobilidade ocorre em grande cidades principalmente devido a falta de infraestruturas necessárias e estacionamentos são um componente essencial na discussão de mobilidade sustentável. A escassez de vagas aumenta o tempo de veículos em trânsito oque causa desperdício de combustível, maior congestionamento e poluição. É preciso, portanto, buscar implementações que utilizam os recursos já existentes nestas cidades. *Smart Parking* possui foco na automatização e melhor gerenciamento de tarefas diárias e funções do estacionamento. Parte desse gerenciamento envolve o preço das vagas que possui influência no funcionamento de um estacionamento. As alterações deste valor podem maximizar a receita ao cobrar valores maiores durante grande demanda ou incentivar o uso das vagas com menores preços ampliando a ocupação quando desejado. Para um processo de reserva de vagas, implementado em um sistema multiagente, é necessário um método de previsão de preço.

Neste trabalho é proposto um sistema de previsão de preços usados em processo de reserva de vagas. Será utilizado um método de aprendizagem de máquina de forma supervisionada sobre dados adquiridos em simulações prévias para a geração de uma árvore de decisão. O algoritmo C4.5 será explorado como método de aprendizado. Ele deve ser implementado como um artefato do ambiente sobre controle de um agente gerente do sistema multiagente. Este gerente terá a função de informar o preço previsto em requisições por vagas futuras e o agente motorista precisa informar a data e hora que pretende usar a vaga.

A árvore criada para este sistema deve conter os fatores que afetaram o valor do preço nas simulações, começando pelo mais influente. A escolha de cada nó segue os seguintes passos: calcular o ganho de informação de cada atributo usando conceito de entropia; criar um nó com a partição de dados do atributo com o maior ganho; repetir para cada subconjunto particionado por este atributo e adicionar como nós filhos. A previsão do preço ocorre pela travessia dessa árvore até o nó folha contendo o preço representativo na situação prevista.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Tendências para o Desenvolvimento de Jogos Sérios na Área de Educação Financeira

Isabel Cristina Torrens[a,c], Simone Nasser Matos[a], Helyane Bronoski Borges[a], Rui Pedro Lopes[b]

[a]*Departamento de Ciência da Computação, Universidade Tecnológica Federal do Paraná, Rua Doutor Washington Subtil Chueire, 330, Ponta Grossa, Paraná, Brasil*
[b]*Centro de Investigação em Digitalização e Robótica Inteligente, Intituto Politécnico de Bragança, Alameda de Santa Apolónia, 253, Bragança, Portugal*
[c]*Autor para correspondência: isabeltorrens.alunos@utfpr.edu.br*

Palavras-chaves: educação financeira, jogos sérios, tendências

A Agenda 2030 foi adotada pela ONU (2015) com o objetivo de desenvolver habilidades, valores e atitudes que assegurem sociedades mais sustentáveis. Um dos temas que contribui para promover o consumo e produção sustentáveis é a Educação Financeira (Moreira e Santos, 2020). Para auxiliar o estudo dessa disciplina, diversos recursos tecnológicos têm sido desenvolvidos, como o uso de jogos sérios, que facilitam a aprendizagem sobre um determinado tema. Buscando identificar como são desenvolvidos para esse tema, foi realizado um mapeamento sistemático que retornou um total de 516 estudos, nos últimos 5 anos. Baseado na metodologia de Kitchenham e Charters (2007), foram definidas questões e strings de buscas, fontes de pesquisa e critérios de seleção. Com a execução do protocolo, a seletiva resultou em 9 artigos com temas relevantes sobre o assunto. O mapeamento mostra que os jogos são, em sua maioria, do tipo simulação e voltados para um público entre 12 e 18 anos. Os principais temas de educação financeira abordados são sobre: consumo consciente, gestão de receitas e despesas e investimentos. Em relação às tecnologias, as mais utilizadas são para plataforma web/mobile, como Java/Web e JQuery. O estudo permitiu identificar como tendências para a criação de jogos sérios destinados ao ensino da educação financeira: i) abordagens para aplicação de jogos sérios com recursos *multiplayer* e *design* de níveis, ii) o uso de inteligência artificial dentro da mecânica do jogo, iii) criação de jogos que abordem temas destinados a alunos do 1º ao 5º ano do ensino fundamental, iv) e desenvolvimento de recursos dentro do jogo que permitam o educador acompanhar o progresso do aluno, para que a disciplina seja trabalhada efetivamente como um tema transversal.

Esta pesquisa é financiada pela Fundação Araucária (88887.630399/2021-00).

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Um método para seleção de atributos em bases de dados de classificação hierárquica multirrótulo

Raimundo Osvaldo Vieira[a,b], Helyane Bronoski Borges[a,c]

[a]*Programa de Pós-Graduação em Ciência da Computação, UTFPR Campus Ponta Grossa, Rua Doutor Washington Subtil Chueire, 330, Ponta Grossa, PR, Brasil*
[b]*Depto. de Computação, IFMA Campus Monte Castelo, Av. Getúlio Vargas, 04, São Luís, MA, Brasil*
[c]*Helyane Bronoski Borges: helyane@utfpr.edu.br*

Palavras-chaves: classificação hierárquica multirrótulo, redução de dimensionalidade, seleção de atributos

O problema de classificação é um dos mais importantes no campo da aprendizagem de máquina e consiste em atribuir um rótulo a um elemento do conjunto de dados alvo da tarefa de classificação. De modo particular, há contextos nos quais um objeto de um conjunto pode ser associado a mais de uma classe e, além disso, é possível, ainda, que as classes mantenham entre si relações taxonômicas. Neste caso, tem-se a classificação hierárquica multirrótulo, que pode ser aplicada para a predição de proteínas em dados de bioinformática, por exemplo. Em geral, este tipo de aplicação possui um grande volume de dados e um número elevado de atributos e rótulos, o que interfere de forma negativa no desempenho do classificador. Para lidar com este problema é necessária a aplicação de técnicas para reduzir a dimensionalidade dos dados sem alterar seu significado intrínseco. A seleção de atributos é uma das abordagens de redução de dimensionalidade em bases de dados, que compreende a escolha dos atributos mais relevantes a partir dos originais. As técnicas de seleção de atributos podem ser classificadas em três categorias: filtro, *wrapper* e embutida. A abordagem filtro corresponde à escolha de um subconjunto de atributos por meio da estimação da qualidade dos atributos com base apenas nos dados. Esta pesquisa tem como objetivo propor um método para seleção de atributos em bases de dados de classificação hierárquica multirrótulo utilizando a abordagem filtro. O método consiste em ranquear os atributos a partir do cálculo do *Fisher Score*, adaptando-se esse cálculo para considerar a natureza hierárquica e multirrótulo da tarefa de classificação. Além disso, o método faz uso do coeficiente de correlação entre os atributos para lidar com redundância entre eles. Em seguida, o conjunto de dados reduzido é avaliado por meio da aplicação de um classificador hierárquico multirrótulo e de sua medida de desempenho. Para testar o método, serão realizados experimentos em bases de dados da Gene Ontology e os resultados serão avaliados com base na taxa de acerto dos classificadores e validados estatisticamente.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Tessellation cover for bounded treewidth graphs

Daniel Fabio Domingues Posner[a,b], Renan Guimarães da Silva[a]

[a] *DCC - Universidade Federal Rural do Rio de Janeiro, Av. Governador Roberto Silveira s/nº. Moquetá – Nova Iguaçu – RJ*
[b] *Corresponding author: posner@ufrrj.br*

Keywords: Graph Theory, Graph Coloring, Computational Complexity

A *tessellation* of a graph G is a partition of its vertex set $V(G)$ into vertex disjoint cliques. A *tessellation cover* of G is a set of tessellations that covers its edge set $E(G)$. The *tessellation cover number* $T(G)$ of a graph G is the size of a minimum tessellation cover of G. The k-TESSELABILITY problem aims to decide whether $T(G) \leq k$ for a graph G and an positive integer k. This problem is $\mathcal{NP}$-complete for any fixed $k \geq 3$, chordal graphs or planar graphs (Abreu et al., The graph tessellation cover number: chromatic bounds, efficient algorithms and hardness, *Theoretical Computer Science* 801, 175-191, 2020). A graph G is a *q-tree* if G is clique of size q or if G is obtained from a q-tree by adding a new vertex adjacent to a clique of size q. A graph G is a *partial q-tree* if it is a subgraph of a q-tree. The *treewidth* of a graph G is the smallest q such that G is a partial q-tree.

The Courcelle's Theorem states that any graph theory problem described in monadic second-order logic (MSOL) can be decided in linear time on bounded treewidth graphs (Courcelle, The monadic second–order theory of graphs I: Recognizable sets of finite graphs, INFORMATION AND COMPUTATION 85, 12–75, 1990). Courcelle's Theorem can also be used with a variation of monadic second-order logic called MSOL2, which allows quantification over the set of edges (cf. Pires et al., Revisando o Teorema de Courcelle, PROCEEDINGS OF LI SBPO, v.2, 107372, 2019). In this work we show that the k-TESSELABILITY problem can be described in monadic second order logic 2 (MSOL_2). As a consequence, we obtain a linear time algorithm to solve k-TESSELABILITY when k is fixed for bounded treewidth graphs.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Coloração de Arestas em Grafos Indiferença

Rafael Porto[a,b], Sheila M. Almeida[a]

[a] *Universidade Tecnológica Federal do Paraná, Departamento Acadêmico de Informática, R. Doutor Washington Subtil Chueire, 330, Ponta Grossa, PR, Brasil*
[b] *rafael.1997@alunos.utfpr.edu.br*

Palavras-chaves: coloração de arestas, grafo de intervalo, grafo indiferença

Uma coloração de arestas é uma atribuição de cores para as arestas de um grafo. A coloração de arestas é própria se arestas que incidem no mesmo vértice têm cores distintas. O menor número de cores para uma coloração de arestas própria de um grafo G é o índice cromático de G, $\chi'(G)$. O Problema da Coloração de Arestas é determinar o índice cromático de um grafo G. O grau máximo de G, $\Delta(G)$, é o maior número de arestas incidentes em um mesmo vértice de G. Por definição, $\chi'(G) \geq \Delta(G)$.

Considere um grafo simples G. Existe um algoritmo polinomial para se obter uma coloração de arestas para G com $\Delta(G)+1$ cores (V. Vizing, *On an estimate of the chromatic class of a p-graph*). Então, $\chi'(G) \in \{\Delta(G), \Delta(G)+1\}$. Se $\chi'(G) = \Delta(G)$, G é classe 1; se $\chi'(G) = \Delta(G)+1$, G é classe 2. Decidir se G é classe 1 é um problema NP-completo, conhecido como Problema da Classificação (I. Holyer, *The NP-Completeness of Edge-Coloring*).

Um grafo de intervalos é um grafo G que representa uma família $\mathcal{F}$ de intervalos da reta real, ou seja, existe uma correspondência biunívoca entre os vértices de G e os intervalos de $\mathcal{F}$; e dois vértices são adjacentes em G se e somente se seus respectivos intervalos se intersectam. Em 1985, D.S. Johnson afirmou que determinar a complexidade computacional do Problema da Classificação restrito aos grafos de intervalos é "possivelmente fácil" (D. Johnson, *The NP-completeness: An Ongoing Guide*). Em 1999, Figueiredo et. al provaram que todo grafo de intervalos G com $\Delta(G)$ ímpar é Classe 1 (C. Figueiredo, J. Meidanis e C. Mello, *Total-chromatic number and chromatic index of dually chordal graphs*).

Um grafo de intervalos é indiferença se representa uma família em que todos os intervalos têm o mesmo tamanho. O Problema da Classificação permanece em aberto mesmo quando restrito aos grafos indiferença. Neste projeto, pretende-se investigar o Problema da Coloração de Arestas e, consequentemente, o Problema da Classificação, para os grafos indiferença.

Dado um grafo G, um $\Delta(G)$-vértice é um vértice de G que tem grau igual a $\Delta(G)$. Denotamos por $d(N(v))$ a soma dos graus dos vértices que são adjacentes a v em G. Seja Ψ a classe dos grafos G tais que todo $\Delta(G)$-vértice v satisfaz $d(N(v)) \leq \Delta(G)^2 - \Delta(G)$. L. Zatesko et. al (Edge-colouring graphs with bounded local degree sums) provaram que todo grafo em Ψ é Classe 1. Este projeto está em fase inicial e pretende caracterizar os grafos indiferença que pertencem a Ψ.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Minimal feedback arc sets containing long paths in few-vertex tournaments

Fábio Botler, Rafael Schneider

Systems Engineering and Computer Science Program
COPPE/UFRJ, Rio de Janeiro, RJ, Brazil

{*fbotler,rschneider*}*@cos.ufrj.br*

Keywords: Minimal feedback arc set, Tournament, Lichiardopol's conjecture

A digraph D is a pair $D = (V, A)$ in which $V = \{v_1, v_2, \ldots, v_n\}$ is a set of *vertices*, and $A = \{e_1, e_2, \ldots, e_m\} \subseteq V \times V$ is a set of *arcs*. The *outdegree* (resp. *indegree*) of v, denoted by $d^+(v)$ (resp. $d^-(v)$), is the number of arcs leaving (resp. coming in) v. We denote by $\delta^+(D)$ the *minimum outdegree* of D, that is, the smallest number of arcs leaving a same vertex. A digraph D is *strongly connected* if for each pair $u, v \in V(D)$ there is a directed path from u to v.

A set of arcs $F \subseteq A(D)$ is a *feedback arc set* (FAS) if the digraph obtained from D by removing the arcs in F contains no directed cycles. Furthermore, a feedback arc set F is *minimal* if there is no feedback arc set F' such that $F' \subsetneq F$. In this work, we explore a conjecture attributed to Lichiardopol (see Sullivan, B. D. "A summary of problems and results related to the Caccetta-Häggkvist conjecture" *arXiv preprint math/0605646*, 2006), which says that every digraph D has a minimal feedback arc set containing a path with length at least $\delta^+(D)$.

Lichiardopol's conjecture would imply the well-known Caccetta-Häggkvist, but, as noted by one of the referees, it seems that a counterexample to it was found by Mader in 2006 (Chudnovsky, M. Seymour, P. and Thomas, R. "The Caccetta-Haggkvist conjecture" Workshop, American Institute of Mathematics, San Jose, California, 2006). Although we could not find Mader's specific construction, after the submission of this abstract, we obtained the following counterexample. For every $k \geq 3$, let D_k be the digraph obtained from k disjoint arborescences of height at least 3 with minimum outdegree at least k and add an arc from each of the leaves to each of the roots. The key observation is that every internal vertex of a path in a FAS has indegree at least 2, but such vertices induce an independent set in D_k.

Although Lichiardopol's Conjecture does not hold in its generality, it is still interesting to explore it in specific classes of digraphs. In this work, we explore Lichiardopol's Conjecture in *tournaments*, which are orientations of complete graphs. More specifically by using a binary search algorithm that uses a special characterization of minimal feedback arc sets, we verified its validity to the 6405 strongly connected tournaments with up to eight vertices.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Flood-it Game on Co-comparability Graphs

Guilherme Quintero Lorenzi[a,b], Sheila Morais de Almeida[a]

[a] *Federal University of Technology – Paraná, Academic Department of Informatics, R. Dr. Washington Subtil Chueire, 330, Ponta Grossa, PR, Brazil*
[b] *Corresponding author: guilhermequintero@alunos.utfpr.edu.br*

Keywords: flood-it, co-comparability graph, computational complexity, flooding game

A flooding game is a game on a pre-colored board where each cell has a color, and in each turn, the player assigns a new color to one of the cells on the board. When a cell receives a new color, it is merged with all its neighbors with the same color. The goal of the game is to obtain a monochromatic board using the minimum number of turns. When the board cells are squares, the game is called Flood-it, and when they are hexagons, the game is called HoneyBee or Mad Virus. There are two versions of these games: in the free-pivot version, the player can choose in each turn what cell he wants to color; in the fixed-pivot version, it is always the same cell that receives a new color.

The flooding games can be generalized for graphs: given a graph whose vertices were previously colored, a new color is assigned to the pivot in each turn and this assignment of color is propagated to every vertex connected to the pivot by a monochromatic path (considering the coloring as when the turn starts). As in the board version, the goal is to obtain a graph whose vertices are all colored the same, using the minimum number of turns.

The fixed-pivot version of the Flood-it game is polynomial-time solvable for interval graphs (Fukui, et. al, On Complexity of Flooding Games on Graphs with Interval Representations, 2012), for the powers of paths P_n^2 and the powers of cycles C_n^2 (Souza, et. al, An algorithmic analysis of Flood-It and Free-Flood-It on graph powers, 2014), and for $2 \times n$ circular grid graphs (Souza, et. al, An algorithmic analysis of Flood-It and Free-Flood-It on graph powers, 2014). However, the free version of the Flood-it game is NP-Complete for interval graphs (Fukui, et. al, On Complexity of Flooding Games on Graphs with Interval Representations, 2012), and it is NP-hard for powers of cycles C_n^2 and $3 \times n$ circular grid graphs (Souza, et. al, An algorithmic analysis of Flood-It and Free-Flood-It on graph powers, 2014).

A co-comparability graph is the intersection graph of curves between two parallel lines. Fleischer and Woeginger (An algorithmic analysis of the Honey-Bee game, 2012) proved that the fixed-pivot Flood-it game is polynomial-time solvable for co-comparability graphs. They present two algorithms to solve that problem: a single source shortest path algorithm and a dynamic programming. In this note, we claim that the recursive definition of their dynamic programming is not well-defined, and we present an example for which it does not finish.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Coloração de Arestas em Grafos Split Proibidos para Comparabilidade e Intervalos

Cintia I. Cararo[a,d], Sheila M. Almeida[a], Cândida N. Silva[b], Glasielly D. Proença[c]

[a] *Universidade Tecnológica Federal do Paraná, Departamento de Informática, R. Dr Washington Subtil Chueire, 330, Ponta Grossa, PR, Brasil*
[b] *Universidade Federal de São Carlos, Centro de Ciências em Gestão e Tecnologia, Departamento de Computação, Rodovia João Leme dos Santos, km 110 - SP-264, Sorocaba, SP, Brasil*
[c] *Universidade Federal de Mato Grosso do Sul, Av. Márcio Lima Nantes, s/n, Coxim, MS, Brasil*
[d] *Autor para correspondência: cintiacararo@alunos.utfpr.edu.br*

Palavras-chaves: coloração de arestas, grafo split, grafo de intervalos, grafo de comparabilidade

O *Problema da Coloração de Arestas* é determinar o menor número de cores k com o qual se pode colorir as arestas de um dado grafo G tal que arestas adjacentes possuam cores distintas. Esse número é chamado de *índice cromático de* G e denotado por $\chi'(G)$. Por definição, $\chi'(G) \geq \Delta(G)$, onde $\Delta(G)$ é o grau máximo do grafo. Vizing mostrou que $\chi'(G) \leq \Delta(G)+1$ para todo grafo simples G (V. G. Vizing, *On an estimate of the chromatic class of a p−graph*). Decidir se $\chi'(G) = \Delta(G)$ é um problema $\mathcal{NP}$-Completo (I. Holyer, *The NP-completeness of edge-coloring*).

Um grafo G é *split* se seu conjunto de vértices pode ser particionado em dois subconjuntos, Q e S, tais que Q é uma clique e S é um conjunto independente. Seja G um grafo split. O Problema da Coloração de Arestas está resolvido quando $\Delta(G)$ é ímpar (B. L. Chen, H. L. Fu e M. T. Ko, *Total chromatic number and chromatic index of split graphs*); quando G tem vértice universal (M. Plantholt, *The chromatic index of graphs with a spanning star*); e em alguns casos de interseção dos split com outras classes de grafos.

As classes de grafos split-comparabilidade e split-intervalos foram caracterizadas por subgrafos induzidos proibidos (S. Foldes e P. L. Hammer, *Split Graphs Having Dilworth Number Two*). O Sol e seu complemento são proibidos para ambas as classes. Para os split-intervalos o terceiro subgrafo proibido é o Sol nascente e para os split-comparabilidade o terceiro proibido é o complemento do Sol nascente. O índice cromático dos grafos split-comparabilidade e split-intervalos é conhecido (L. G. S. Gonzaga, J. B. S. Cruz, C. N. Silva e S. M. Almeida, *The Overfull Conjecture on Split-Comparability and Split-Interval Graphs*). Assim, pretendemos investigar o Problema da Coloração de Arestas nos grafos split com grau máximo par que contém subgrafos proibidos para ambas as classes, ou seja, que contém o Sol, ou seu complemento, ou contém o Sol nascente e seu complemento.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Deep Learning aplicado à estimativa de profundidade em imagens estéreo multivista

Sérgio M. M. de Oliveira[a,b,d], Erikson F. de Morais[a,b], Marcella S. R. Martins[a,c]

[a] *Universidade Tecnológica Federal do Paraná, Ponta Grossa, Paraná, Brasil*
[b] *Departamento Acadêmico de Informática - DAINF*
[c] *Departamento Acadêmico de Eletrônica - DAELE*
[d] *Autor para correspondência: sergiomurilo@alunos.utfpr.edu.br*

Palavras-chaves: visão estéreo multivista, automação veicular, aprendizado profundo

Estimar a distância entre o ponto de observação e os objetos no ambiente é parte crucial no desenvolvimento de diversas aplicações robóticas.

Atualmente, o LiDAR (*Light Detection and Ranging*) é um dos principais sensores responsáveis pela estimativa de profundidade. Com margem de erro de 2 centímetros a 200 metros do alvo nos modelos mais avançados, fornece informações precisas quanto à profundidade dos objetos nas cenas observadas. Em contrapartida, o alto custo de produção inviabiliza a utilização em larga escala do LiDAR, pois a necessária redundância de sensores em veículos autônomos, por exemplo, os torna comercialmente inviáveis.

Imagens digitais, monoculares e binoculares (captadas a partir de duas câmeras observando a mesma a cena de pontos de vista ligeiramente distintos), apresentam indicadores visuais (*depth cues*). A aplicação de técnicas de Visão Computacional no processamento de tais imagens permite inferir a profundidade da cena observada. Ainda que possuam alta resolução, câmeras digitais são mais acessíveis e baratas em comparação ao LiDAR. Por outro lado, o custo computacional, diretamente proporcional à resolução das imagens processadas, e a baixa precisão em comparação ao LiDAR dificultam a utilização de tais técnicas em aplicações comerciais.

Observando tal cenário, o presente trabalho aplica *Deep Learning* na obtenção do mapa de profundidade da cena observada. Os experimentos atuais, realizados com imagens estéreo provenientes do *DrivingStereo*[1] *Dataset* atingiram 54 % de acurácia, medida a partir da comparação entre o mapa de profundidade real (fornecido pelo *dataset*) e o mapa gerado pelo modelo. O modelo atual consiste em uma Rede Neural Siamesa com 2 sub-redes compostas por camadas convolucionais, responsáveis por processar a imagem esquerda e direita simultaneamente. O resultado passa por uma camada de concatenação e é utilizado como semente para uma Rede GAN (*Generative Adversarial Network*), que gera o mapa de profundidade da cena observada.

[1] Yang, Guorun, et al. "Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Definição de pesos aos atributos em uma base de dados sem domínio definido no Raciocínio Baseado em Casos

Jonathan William Hrenechen[b,c], André Koscianski[a,b], André Pinz Borges[a,b]

[a]*Departamento Acadêmico de Informática, Pós-Graduação em Ciência da Computação, Av. Monteiro Lobato, Km 4, s/n, Ponta Grossa - Paraná - Brasil*
[b]*Universidade Tecnológica Federal do Paraná, Av. Monteiro Lobato, Km 4, s/n, Ponta Grossa - Paraná - Brasil*
[c]*Autor para correspondência: hrenechen@gmail.com*

Palavras-chaves: RBC, peso de atributos, data-driven

O método de Raciocínio Baseado em casos (RBC), utiliza a experiência passada para solucionar um novo problema, buscando a solução mais próxima dentro de uma base de casos. Cada caso é composto por um conjunto de atributos e uma categoria. O metodo de RBC depende de um especialista no domínio que vai ser aplicado para avaliar quais são as informações relevantes para gerar uma solução, neste quesito a abordagem 'data-driven', explora a falta deste especialista ou mesmo um especialista com pouca competência no domínio apresentado.

Na abordagem data-driven é criado um placar com a pontuação dos atributos, em que apenas uma porcentagem delas será utilizada. O placar é ordenado por relevância e para isso podem-ser aplicados métodos como Information Gain, Gain Ratio, Gini, Chi2, ReliefF.

O cálculo de pontuação é realizado dentro do intervalo [0, 1], sendo que cada dimensão (atributo) da base de dados pode receber um peso diferente. Mudanças nos pesos resultam em uma nova base de dados, sua ordenação é relacionada ao peso de cada atributo, e a porcentagem de utilização irá excluir os atributos com menor pontuação, exceto nos casos que 100% da base seja utilizada.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Cálculo de raízes quadradas de grafos utilizando um resolvedor pseudo-booleano

Pietro Cavassin[a,b], Aleffer Rocha[a], Renato Carmo[a]

[a] *Departamento de Informática, Universidade Federal do Paraná, Rua Cel. Francisco Heráclito dos Santos, 100 – Centro Politécnico, Curitiba, Paraná, Brasil*
[b] *Autor para correspondência: pietro.polinari@gmail.com*

Palavras-chaves: raiz quadrada de grafos, programa pseudo-booleano, satisfazibilidade

O quadrado de um grafo H é o grafo H^2 dado por $V(H^2) = V(H)$ e $E(H^2) = \{uv\colon d_H(u,v) \leq 2\}$. Dizemos então que H é uma raiz quadrada de H^2. Existem grafos com nenhuma, uma ou mais de uma raiz quadrada. O problema da Raiz Quadrada consiste em determinar se um grafo dado tem raiz quadrada. Trata-se de um problema $\mathcal{NP}$-completo (R. Motwani e M. Sudan: *Computing roots of graphs is hard*, 1994) que pode ser reduzido ao problema de satisfazibilidade booleana (P. L. Hammer e S. Rudeanu: *Pseudo-boolean programming*, 1969). Se G e H são tais que H é uma raiz quadrada de G, então $M_G = M_H + M_H^2$, onde M_G e M_H são suas respectivas matrizes de adjacência e as operações são booleanas. Essa igualdade define uma instância do problema de Satisfazibilidade Booleana: cada restrição corresponde a uma aresta de G e cada variável corresponde a uma aresta de H. Pode-se assim utilizar um resolvedor pseudo-booleano para computar raízes quadradas de G. Neste trabalho desejamos avaliar até que ponto é vantajoso pré-processar o programa pseudo-booleano antes de passá-lo ao resolvedor. Para tanto utilizaremos o Solving Constraint Integer Programs (SCIP, `https://www.scipopt.org/`). O pré-processamento que pretendemos realizar consiste em diminuir o número de variáveis e equações quando o valor de uma variável já é pré-determinado: se uma aresta não existe no grafo de entrada, ela também não pode existir numa sua raiz quadrada e assim o valor da variável correspondente no programa pseudo-booleano tem que ser zero. Investigar a possibilidade de haver outros pré-processamentos que possam ser realizados a fim de diminuir o tempo de pré-processamento do resolvedor será o próximo passo.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Configuração de Parâmetro no Algoritmo de Agrupamento de Dados baseado em Formigas

Carina Moreira Costa[a,c], Rosangela Villwock[b,c]

[a] *Universidade de Trier, Trier, Alemanha*
[b] *Universidade Estadual do Oeste do Paraná, Cascavel, Paraná, Brasil*
[c] *carinamath5@gmail.com, rosangela.villwock@unioeste.br*

Resumo

O método de Agrupamento de Dados baseado em Colônia de Formigas é um método bioinspirado no comportamento de colônias de formigas reais, o qual ainda precisa de muita investigação para se tornar uma ferramenta madura para mineração de dados. Este algoritmo possui a grande vantagem de não solicitar o número de grupos e uma desvantagem, a dificuldade de configuração de outros parâmetros. Nesse sentido, o objetivo deste trabalho foi melhorar a eficácia do algoritmo, propondo uma estratégia para configuração de um destes parâmetros, considerado o parâmetro mais difícil de ser calibrado. Para isso, foram realizados experimentos utilizando bases de dados reais e públicas, e com análises da matriz de dissimilaridade foi possível estabelecer um procedimento para definição do parâmetro. Ao final deste estudo, observou-se a utilidade e validade da estratégia proposta, pois se trata de uma forma simples de se definir o parâmetro e o desempenho do algoritmo foi razoável se comparado a execuções nas quais não se sabe como defini-lo ou nas quais se usa uma configuração padrão.

Abstract

The Ant Colony based Data Clustering method is a method bio-inspired on the behavior of real ant colonies, which still requires a lot of research to become a mature tool for data mining. This algorithm has the great advantage of not requiring the number of clusters and a disadvantage, the difficulty of setting other parameters. Therefore, the goal of this work was to improve the effectiveness of the algorithm by proposing a strategy to set up one of those parameters, considered to be the most difficult parameter to be calibrated. To this end, experiments were conducted using real-world and public data sets, and with analyses of the dissimilarity matrix it was possible to establish a procedure for defining the parameter. With this study, the usefulness and validity of the proposed strategy was verified, since it is a simple way to define the parameter and the performance of the algorithm was reasonable when compared with the experiments in which it is not known how to define the parameter or in which a default configuration is used.

Palavras-chaves: Mineração de Dados, Metaheurística, Matriz de dissimilaridade

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

1. Introdução

O algoritmo de Agrupamento de Dados baseado em Colônia de Formigas, diferente da maioria dos algoritmos de agrupamento de dados, possui a vantagem de não requerer como parâmetro o número de grupos [1] . Entretanto, este algoritmo possui outros parâmetros que são difíceis de serem determinados e alguns dependem da base de dados. O ajuste dos parâmetros é crucial para obter um bom agrupamento [2].

Estudos envolvendo este algoritmo são relevantes dado que trata-se de uma metaheurística relativamente nova e que tem recebido atenção especial, principalmente porque ainda exige muita investigação para melhorar seu desempenho, estabilidade e outras características, que fariam de tal algoritmo uma ferramenta madura para mineração de dados [3].

De acordo com Lauro [1], o problema da convergência do algoritmo é muito citado na literatura e já existem algumas propostas para resolvê-lo. Porém, estas propostas fogem um pouco do paradigma inicial do algoritmo, que tem como virtude a simplicidade. Conforme Faria, Stephany e Becceneri [2], a definição e calibração dos parâmetros do algoritmo para obter uma execução com bom desempenho têm dificultado e afastado os usuários deste.

Assim, o objetivo deste trabalho foi buscar a melhoria da eficácia do algoritmo de Agrupamento de Dados baseado em Colônia de Formigas por meio de um melhor ajuste de um de seus parâmetros.

2. Agrupamento de Dados baseado em Colônia de Formigas

No Agrupamento de Dados baseado em Colônia de Formigas proposto por Deneubourg et al. [4] as formigas foram representadas como agentes simples que se moviam aleatoriamente em uma grade quadrada. Os padrões foram dispersos nesta grade e poderiam ser carregados, transportados e descarregados pelas formigas. Estas operações são baseadas na similaridade e na densidade dos padrões numa vizinhança local. Após um número estipulado de iterações, se obtém uma separação espacial dos padrões na grade e os grupos formados tendem a formar regiões densas na grade.

Na proposta de Deneubourg et al. [4], as decisões de carregar e descarregar padrões são tomadas pelas probabilidades P_{pick} e P_{drop}, dadas pelas Equações (1) e (2). Nestas equações, $f(i)$ é uma estimativa da fração de padrões na vizinhança que são semelhantes ao padrão atual i da formiga e k_p e k_d são constantes reais. Os parâmetros k_p e k_d determinam a influência que a função densidade $f(i)$ exerce no cálculo da probabilidade de um agente carregar ou descarregar um objeto em determinada posição. A função de vizinhança $f(i)$ tem portanto grande influência nas decisões de carregar e descarregar padrões.

Lumer e Faieta [5] incluíram modificações ao modelo que permitiram a manipulação de dados numéricos com o uso de diferentes tipos de objetos e melhoraram a qualidade da solução e o tempo de convergência do algoritmo. Uma das mudanças é a da Equação (3), apresentada a seguir. Nesta equação, $d(i, j)$ é uma função de dissimilaridade entre padrões i e j pertencente ao intervalo $[0, 1]$, o parâmetro $\alpha \in (0, 1]$ vai escalar a dissimilaridade e L é a vizinhança local de tamanho igual a σ^2. Usualmente, $\sigma^2 \in \{9, 25\}$.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Handl, Knowles e Dorigo [6] também modificaram o algoritmo, especificamente o cálculo das probabilidades de carregar e descarregar padrões, bem como incluíram uma restrição na função de vizinhança dada pela Equação (4). Com esta restrição dissimilaridades elevadas são penalizadas, o que segundo os autores, melhora a separação espacial entre os grupos e acelera o tempo de execução do algoritmo.

$$P_{\text{pick}} = \left(\frac{k_p}{k_p + f(i)} \right)^2 \tag{1}$$

$$P_{\text{drop}} = \left(\frac{f(i)}{k_d + f(i)} \right)^2 \tag{2}$$

$$f(i) = max \left\{ 0, \ \frac{1}{\sigma^2} \sum_{j \in L} \left[1 - \frac{d(i,j)}{\alpha} \right] \right\} \tag{3}$$

$$\bar{f}(i) = \begin{cases} \frac{1}{\sigma^2} \sum_{j \in L} \left[1 - \frac{d(i,j)}{\alpha} \right], & \text{se} \quad \forall j \ \left(1 - \frac{d(i,j)}{\alpha} \right) > 0 \\ 0, & \text{caso contrário.} \end{cases} \tag{4}$$

3. Metodologia

A implementação do método utilizado neste trabalho foi a de Boscarioli et al. [7], na ferramenta YADMT - *Yet Another Data Mining Tool*. Maiores detalhes da implementação podem ser vistos em Costa [8]. O parâmetro σ foi escolhido empiricamente, para cada base de dados, dentre valores encontrados na literatura. O parâmetro escolhido para estudo neste trabalho foi o limitante de dissimilaridade, α, que vai "escalar" ou "limitar" a dissimilaridade na função de vizinhança $f(i)$. O funcionamento do algoritmo é crucialmente dependente deste parâmetro, sendo esse o parâmetro mais difícil de ser calibrado. A escolha de um valor muito pequeno para α impede a formação de grupos na grade, por outro lado, a escolha de um valor muito grande resulta na fusão de grupos [6].

A ideia foi usar a matriz de dissimilaridade normalizada no intervalo $[0, 1]$ e a restrição na função de vizinhança proposta por Handl, Knowles e Dorigo [6], dada pela Equação (4). Esta matriz apresenta as distâncias entre todos os padrões, e a partir da análise desta é possível estimar o parâmetro α. Cada base de dados terá um valor para α.

Para a definição deste parâmetro, foram calculadas as porcentagens de distâncias entre padrões maiores do que um certo valor (candidato a ser o valor de α), observado no intervalo [0,1; 0,9] considerando incremento de 0,01. O valor de α_{inicial} foi definido como o valor (a distância) que melhor diferenciava os dados na matriz de dissimilaridade (0,3 para a base de dados IRIS pois aproximadamente 56% das distâncias entre padrões são maiores que 0,3). Assim, de acordo com a restrição que penaliza dissimilaridades elevadas na Equação (4), temos que $d(i, j) > \alpha_{\text{inicial}}$ para 56% das distâncias e, portanto, os padrões que atendem a esta condição são considerados dissimilares, o que resulta em uma separação boa dos dados

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

já no início do algoritmo. Para uma estimativa do parâmetro α_{max} foi definido um intervalo de valores observando a matriz de dissimilaridade. O limite inferior foi definido como o valor cuja porcentagem de distâncias entre padrões maiores que ele era aproximadamente 40% e o limite superior foi definido como o valor cuja porcentagem de distâncias entre padrões maiores que ele era aproximadamente 20%. Para uma estimativa do parâmetro α_{controle} foi definido um intervalo de valores. O limite inferior foi definido como o valor cuja porcentagem de distâncias entre padrões maiores que ele era aproximadamente 80% e o limite superior foi definido como o valor cuja porcentagem de distâncias entre padrões maiores que ele era aproximadamente 70%. O parâmetro α_{min} foi definido com o mesmo valor de α_{inicial}. Os percentuais acima foram definidos previamente.

4. Resultados e Discussão

Para a escolha do parâmetro α adotou-se um critério para a definição de um intervalo de busca e, após isso, os valores foram definidos empiricamente por meio de experimentos. Os resultados obtidos considerando-se as configurações definidas para cada base de dados podem ser vistos em Costa [8].

A Tabela 1 apresenta a comparação das médias das medidas de avaliação (Índice Aleatório - R, Medida F e porcentagem de acerto) obtidas neste trabalho com as que foram obtidas aplicando o algoritmo utilizando a configuração padrão da ferramenta utilizada (YADMT). Comparando os resultados obtidos observa-se a superioridade dos resultados deste trabalho, revelando a efiácia da estratégia proposta.

Tabela 1: Comparação dos resultados médios da aplicação do algoritmo com a estratégia proposta e resultados obtidos utilizando-se a configuração padrão da ferramenta.

Conjunto	Métrica	Estratégia proposta	Configuração padrão
IRIS	R	0,790	0,581
	F	0,785	0,489
	% de acerto	78,3	45,2
WINE	R	0,785	0,555
	F	0,799	0,477
	% de acerto	78,4	46,2
ZOO	R	0,830	0,746
	F	0,672	0,493
	% de acerto	56,6	38,6
PIMA	R	0,552	0,509
	F	0,663	0,585
	% de acerto	66,1	55,5

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

5. Considerações Finais

O algoritmo de Agrupamento de Dados baseado em Colônia de Formigas tem despertado interesse dos pesquisadores em mineração de dados pelos resultados promissores que oferece. Entretanto, a dificuldade de configuração e calibração dos parâmetros afasta alguns usuários deste. A estratégia proposta neste trabalho se mostrou válida e util, pois é uma maneira simples de se definir o parâmetro e o desempenho do algoritmo foi razoável se comparado a execuções nas quais não se sabe como definir ou nas quais se tem uma configuração padrão oferecida pela ferramenta em uso.

Referências

[1] LAURO, A. L.. Agrupamento de Dados Utilizando Algoritmo de Colônia de Formigas. Dissertação (Mestrado) - Programa de Pós-graduação em Engenharia Civil, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2008.

[2] FARIA, G. F.; STEPHANY, S.; BECCENERI, J. C.. Uma Nova Estratégia Acoplada De Inicialização e Ajuste adaptativo do Parâmetro de Similaridade num Algoritmo de Agrupamento Baseado em colônia de formigas. 10a Conferência Brasileira de Dinâmica, Controle e Aplicações, São Paulo, p.415-419, ago. 2011.

[3] BORYCZKA, U.. Finding groups in data: Cluster analysis with ants. Applied Soft Computing 9, 61-70, 2009.

[4] DENEUBOURG, J. L.; GOSS, S.; FRANKS, N.; SENDOVA-FRANKS, A.; DETRAIN, C.; CHRÉTIEN, L. The dynamics of collective sorting: Robot-like ants and ant-like robots. Proceedings of the First International Conference on Simulation of Adaptive Behaviour: From Animals to Animats 1, 356–365, 1991.

[5] LUMER, E.; FAIETA, B. Diversity and adaptation in populations of clustering ants. Proceedings of the Third International Conference on Simulation of Adaptive Behaviour: From Animals to Animats 3, 501–508, 1994.

[6] HANDL, J.; KNOWLES, J.; DORIGO, M. Ant-Based Clustering and Topographic Mapping. Artificial Life v. 12, 35-61, 2006.

[7] BOSCARIOLI, C.; TEIXEIRA, M. F.; VILLWOCK, R.; FAINO, T. M. O módulo de agrupamento de dados da ferramenta YADMT. In: V EPAC - Encontro Paranaense de Computação. V EPAC - Encontro Paranaense de Computação, Cascavel, 2013.

[8] COSTA, C. M. Algoritmo de Agrupamento de Dados Baseado em Colônia de Formigas. Monografia — Universidade Estadual do Oeste do Paraná, Cascavel, 2016.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

A Deductive-Formal Derivation for the Preferential Attachment Metric for Link Prediction

Alane M. de Lima[a], Murilo V. G. da Silva[a], André L. Vignatti[a,b]

[a]*Department of Computer Science, Federal University of Parana, Curitiba, Brazil*
[b]*Corresponding author: vignatti@inf.ufpr.br*

Abstract

We present an analytical derivation of the *preferential attachment metric* to predict social ties in complex networks. This metric was originally proposed by Newman in 2001 and by Barabási et al. in 2002 and it was obtained by empirical means. We propose in this paper an analysis based on a *deductive-formal reasoning*, giving the metric a formal theoretical basis. In our analysis we use two random graph models for power-law graphs. We show that in these models, by using formal reasoning, we can derive the preferential attachment metric proposed by Newman and Barabási et al.

Keywords: link prediction, preferential attachment, formal analysis.

1. Introduction

For a given *complex network* and a pair of its non-adjacent vertices, the link prediction problem asks for the probability that these vertices may become connected by an edge in a near future. Even though this problem has been studied since early 2000's [1], it still remains challenging and has been focus of recent research on the area of complex networks [2, 3]. As pointed by Liben and Kleinberg [1], most of the existing link prediction methods are based on network metrics (e.g. common neighbors, Jaccard measure, Katz centrality). In this work, we are particularly interested in methods based on one of these type of metrics. More precisely, we deal with the *preferential attachment* (PA) metric [1, 4, 5], which is suitable for networks modeled by graphs that have a degree distribution following a power-law function, which are referred here as power-law graphs. Power-law graphs appear in many contexts, including technological, biological, and social networks, so the subject has attracted attention lately from many authors [2, 3, 6]. When first proposed [4, 5], the PA metric was obtained empirically, from experiments on real networks. Indeed, as observed by Liben and Kleinberg [1] in the context of authorship and co-authorship in research publications, where vertices represent researchers and an edge between two vertices x and y indicates that x and y have published a paper together, *"Newman [5] and Barabási et al. [4] have further proposed, on the basis of empirical evidence, that the probability of co-authorship of x and y is correlated with the product of the number of collaborators of x and y"*. This raises an issue: from the

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

point of view of formal sciences (such as logic, mathematics, statistics, theoretical computer science, etc.), the lack of a formal proof is a stumbling block for further results regarding the PA metric.

In this paper we show a deductive-formal derivation of the PA metric. More specifically, we start with three random graph models (as the premises of the formal system), and then we give an analytical derivation for the PA metric in these models. We conclude that the result is considered epistemologically stronger since the result holds both in the formal and empirical paradigms.

2. Preliminaries

Reasoning Methods: deductive reasoning is the process of reasoning from one or more statements (premises) to reach a logically certain conclusion. If all premises are true and the rules of deductive logic are followed, then the conclusion is necessarily true [7]. Inductive reasoning (as opposed to deductive) is a reasoning in which the premises are seen as providers of strong evidence for the truth of the conclusion [7]. Formal sciences are constructed on formal-deductive arguments (theorems), and inductive reasoning is not accepted. Thus, in the context of a formal epistemology, a PA metric cannot be used unless a proper derivation (i.e., a proof) is provided.

Models: A model is a simplified abstract view of a complex reality. Ideally the formal model should not produce theoretical consequences that are contrary to what is found in reality [8, 9]. Typically, however, the model is an approximation, since a complete and precise representation of the reality may be impossible [10]. Thus, it is commonplace that several different models are considered to represent the same reality. In our context, the models are the initial premises of the deductive reasoning process.

Power-law Graphs: The empirical study of large real-world networks in the late 1990's and early 2000's [11, 12] showed that the number of nodes of a given degree i is proportional to $i^{-\beta}$, where $\beta > 0$. Such degree distribution is called *power law*. This discovery was soon followed by works describing models for those networks, which involve the use of random graphs with power-law degree distribution, referred here as *power-law graphs*.

Preferential Attachment (PA) Metric: This metric was proposed by Newman [5] and Barabási et al. [4] independently, through experiments. Given a pair of vertices (u, w), the PA metric is defined by $\mathrm{PA}(u, w) = \frac{d(u)d(w)}{\sum_v d(v)}$, where $d(v)$ is the degree of vertex v of a given graph.

So, starting from the power-law graph models as the premises, we provide an analysis that has all steps based on deductive reasoning and, in the end, we give a definition of the PA metric.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

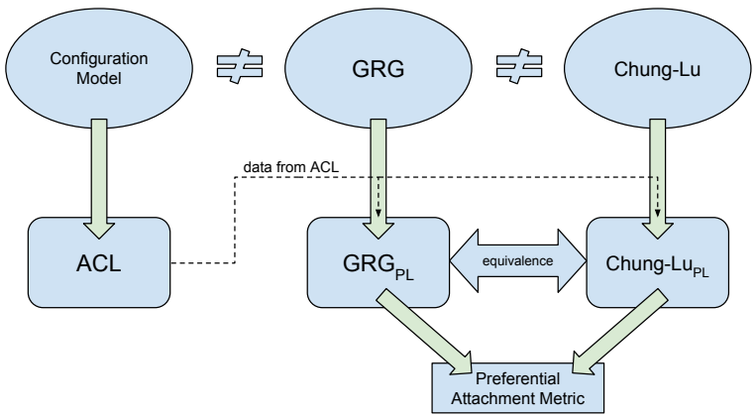


Figure 1: The complete path of the analysis.

3. Analysis

Our analysis is an interplay between several models, with theorems supporting the intended result (see Figure 1 for a scheme). As common practice in deductive reasoning, we take the models as premises, from where the theorems are then derived. Next, we briefly describe the models that we used (for an in-depth exposition, see [13]). Throughout this paper, we use $\approx$ to denote an *asymptotic approximation*.

Configuration Model [14]: this is a random graph model suitable when we need certain degree distribution of the graph. The model uses an approach where edge connections are probabilistic, but the list of vertex degrees (and therefore the number of edges) is fixed. Let $d(v)$ be the degree of a vertex $v = 1, \ldots, n$. The idea is to create a random graph in the following way. For each vertex v, let it have $d(v)$ "half edges" incident to it. Now pick a random perfect matching between the half edges pairing them and connecting them together. As a result we obtain a graph such that each vertex has exactly the desired degree $d(v)$. In many contexts, proving results in this model is not an easy task, since the probabilistic events carry dependencies, and that makes it very hard to obtain useful results, in particular for algorithm design.

$ACL(\alpha, \beta)$ Model [15]: the model is obtained from the configuration model using a specific degree distribution for a power-law function. The model depends on input parameters α and β where, roughly speaking, α corresponds to the logarithm of the graph size, and β is the power-law constant (weakly) related to the graph density. Let y_i be the number of vertices with degree i. In this model, by definition, $y_i = \lfloor \frac{e^\alpha}{i^\beta} \rfloor$. Notice that the maximum degree of the graph is $\lfloor e^{\alpha/\beta} \rfloor$ (which is the case where $y_i = 1$). Also, we can ignore rounding [13]. Assuming[1] $\beta > 2$, the number of vertices is $\sum_{i=1}^{e^{\frac{\alpha}{\beta}}} \frac{e^\alpha}{i^\beta} \approx \zeta(\beta) e^\alpha$ and the number of edges is

[1]This is for the purpose of easing the exposure, as [13] deals with other cases for β.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

$\frac{1}{2}\sum_{i=1}^{e^{\frac{\alpha}{\beta}}} i\frac{e^{\alpha}}{i^{\beta}} \approx \frac{1}{2}\zeta(\beta-1)e^{\alpha}$, where $\zeta(t) = \sum_{n=1}^{\infty}\frac{1}{n^t}$ is the *Riemann zeta function*.

Generalized Random Graph (GRG) Model [16]: in this model edges are added independently, and weights on the vertices are used to generate arbitrary distribution of the degrees, including the power-law distribution. Starting with a vertex set $V = \{1, 2, ..., |V|\}$, let each $v \in V$ have weight w_v. Let w be an array with entries $w_1, ..., w_{|V|}$. In the edge set E, every edge uz is created independently with probability $\mathbb{P}(uz \in E) = \frac{w_u w_z}{\ell_n + w_u w_z}$, where $\ell_n = \sum_{v\in V} w_v$. Note that the degree distribution depends on w. To obtain a power-law distribution, we build w using the degree distribution from the ACL model, which is a power-law distribution. However, note that while the ACL model is a specification of the *configuration model*, here we are providing a specification of the GRG model. In Section 3.1 we give more details.

Chung–Lu Model [17, 18]: this model is similar to the GRG model, the difference relying on the probability $\mathbb{P}(uz \in E) = \frac{w_u w_z}{\ell_n}$, where again $\ell_n = \sum_{v\in V} w_v$. For some values of w_v, there is hardly any difference between $\mathbb{P}(uz \in E) = \frac{w_u w_z}{\ell_n + w_u w_z}$ and $\mathbb{P}(uz \in E) = \frac{w_u w_z}{\ell_n}$. In fact, as shown in [13] (Theorem 6.19), when $\sum_{v\in V} w_v^3 = o(n^{3/2})$, GRG and Chung–Lu models are equivalent, but this general condition is not always met for power-law graphs. In Section 3.1, we show how these two models are related.

3.1. Deriving the Metric

Our goal in this section is to derive a formula for the *probability* of an edge appearing between two vertices that is consistent with the empirical definition of the metric.

Due to dependency between events in the configuration model, it is unclear how to obtain a probability for the link prediction directly for the ACL model when it is used as a power-law random graph model. However, since some parameters in the ACL model (e.g., y_i, α, and β) are convenient when dealing with power-law graphs, we use them as mathematical tools to specify the parameters for the GRG and the Chung-Lu models. In the GRG model, we set the weights w so that y_i of its entries are equal to i, for $i = 1, ..., \Delta$, where $\Delta = \lfloor e^{\alpha/\beta} \rfloor$ (note that, similarly, in the ACL model there are y_j vertices of a given fixed degree j). Now, the specific GRG model with power-law parameters creates an edge uz independently with probability $\mathbb{P}(uz \in E) = \frac{w_u w_z}{\ell_n + w_u w_z} \approx \frac{w_u w_z}{\zeta(\beta-1)e^{\alpha} + w_u w_z}$, where $\ell_n = \sum_{v\in V} w_v$.

Lemma 3.1 (Vignatti and da Silva (2016) [6], Lemma 2.1)**.** *Let* $w_u, w_v \in \{1, \ldots, e^{\frac{\alpha}{\beta}}\}$. *Then* $e^{\alpha}\zeta(\beta-1) + w_u w_v \approx e^{\alpha}\zeta(\beta-1)$.

By Lemma 3.1, $\mathbb{P}(uz \in E) \approx \frac{w_u w_z}{\zeta(\beta-1)e^{\alpha}} \approx \frac{w_u w_z}{\ell_n}$, which is the probability of creating an edge in the Chung–Lu model. The results then hold on both GRG and Chung–Lu models[2].

Theorem 3.2. *Let* $v \in V$ *such that* $w_v = k$. *Then* $\mathbb{E}[d(v)] \approx k$, *where* $d(v)$ *is the degree of* v.

[2]The Chung–Lu model can be shown to be asymptotically equivalent to the GRG model under conditions (see [13], Section 6.8.1) that are not necessarily true in our choice of w.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Proof. Let W_i be the set of vertices with weight i, i.e., $W_i = \{v \in V \mid w_v = i\}$ in GRG or Chung–Lu models following the ACL power-law distribution in the assignment of the weights. Let $d_i(v)$ be the number of neighbors of v in W_i.

Note that $d_i(v)$ is a binomial random variable with parameters $n = |W_i|$ and $p = \frac{ik}{e^\alpha \zeta(\beta-1)}$. Therefore,

$$\mathbb{E}[d_i(v)] = |W_i| p_{ik} = \left(\frac{e^\alpha}{i^\beta}\right)\left(\frac{ik}{e^\alpha \zeta(\beta-1)}\right) = \frac{k}{i^{\beta-1}\zeta(\beta-1)},$$

which leads to

$$\mathbb{E}[d(v)] = \sum_i \mathbb{E}[d_i(v)] = \sum_i \frac{k}{i^{\beta-1}\zeta(\beta-1)} \approx \frac{k}{\zeta(\beta-1)}\zeta(\beta-1) = k.$$

The first equality is by linearity of expectation and $d(v) = \sum_i d_i(v)$. □

By Theorem 3.2, our model has power-law expected distribution, since it has $\frac{e^\alpha}{i^\beta}$ vertices of expected degree i, for each $1 \leq i \leq \Delta$. Next, we show that in these models the theoretical and empirical definitions of the PA metric are consistent to each other (i.e., it depends on the degrees of the vertices).

Theorem 3.3 (Preferential Attachment Metric)**.** *Consider $u, z \in V$ in the models GRG or Chung–Lu following the ACL power-law distribution in the assignment of the weights. Then $\mathbb{P}(uz \in E) \approx \frac{d(u)d(z)}{|E|}$, where $d(u)$ and $d(z)$ are the degrees of u and z, respectively.*

Proof. $\mathbb{P}(uz \in E) = \frac{w_u w_z}{\ell_n + w_u w_z} \approx \frac{w_u w_z}{\ell_n} \approx \frac{d(u)d(z)}{|E|}$, where the first equality is from the definition of the GRG model, the first approximation uses Lemma 3.1, and the second approximation uses Theorem 3.2, noting that $\ell_n \approx |E|$. □

4. Conclusion and Future Work

In this work we showed that the PA metric empirically proposed in [5, 4] is the same as the one we derived through our analysis using deductive reasoning. This result gives the metric a formal theoretical basis, so that it can be investigated in the framework of formal sciences. Since many link prediction metrics originate from the inductive paradigm of reasoning, we suggest using deductive analysis on such metrics as future work.

References

[1] D. Liben-Nowell, J. Kleinberg, The link-prediction problem for social networks, J. Am. Soc. Inf. Sci. Technol. 58 (2007) 1019–1031.

[2] C. Lee, M. Pham, N. Kim, M. K. Jeong, D. K. Lin, W. A. Chavalitwongse, A novel link prediction approach for scale-free networks, in: Proc. of the 23rd Int. Conf. on World Wide Web, 2014, pp. 1333–1338.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

[3] Z. Wang, Y. Wu, Q. Li, F. Jin, W. Xiong, Link prediction based on hyperbolic mapping with community structure for complex networks, Physica A: Stat. Mech. and its Apls. 450 (2016) 609–623.

[4] A. L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, T. Vicsek, Evolution of the social network of scientific collaborations, Physica A: Statistical Mechanics and its Applications 311 (2002) 590–614.

[5] M. Newman, Clustering and preferential attachment in growing networks, Phys. Rev. E 64 (2001) 025102.

[6] A. Vignatti, M. da Silva, Minimum vertex cover in generalized random graphs with power law degree distribution, Theoretical Computer Science 647 (2016) 101–111.

[7] I. Copi, C. Cohen, D. Flage, Essentials of Logic, Pearson Prentice Hall, 2007.

[8] H. Freudenthal, The Concept and the Role of the Model in Mathematics and Natural and Social Sciences, Springer Netherlands, 1961.

[9] T. Ritchey, Outline for a morphology of modelling methods, Acta Morphologica Generalis 1 (2012).

[10] G. E. P. Box, N. R. Draper, Empirical Model-building and Response Surface, John Wiley & Sons, Inc., 1986.

[11] M. Faloutsos, P. Faloutsos, C. Faloutsos, On power-law relationships of the internet topology, SIGCOMM Comput.Com.Rev. 29 (1999) 251–262.

[12] J. Kleinberg, S. Laurence, The structure of the web, Science 294 (5548) (2001) 1849–1850.

[13] R. Hofstad, Random Graphs and Complex Networks, Cambridge, 2016.

[14] E. A. Bender, The asymptotic number of labeled graphs with given degree sequences, Journal of Combinatorial Theory A. 24 (3) (1978) 296–307.

[15] W. Aiello, F. Chung, L. Lu, A random graph model for power law graphs, Experimental Mathematics 10 (2001) 53–66.

[16] T. Britton, M. Deijfen, A. Martin-Lof, Generating simple random graphs with prescribed degree distribution, Journal of Statistical Physics 124 (2006) 1377–1397.

[17] F. Chung, L. Lu, Connected components in random graphs with given expected degree sequences, Annals of Combinatorics 6 (2) (2002) 125–145.

[18] F. Chung, L. Lu, The average distance in a random graph with given expected degrees, PNAS 99 (25) (2002) 15879–15882.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Criminalidade e Pandemia: Uma análise do Impacto da COVID-19 na Incidência de Crimes em Curitiba

Patriky Edeovan Galvão Mirkoski

Universidade Tecnológica Federal do Paraná, Av. Sete de Setembro, 3165, Curitiba, Paraná, Brasil

patriky@alunos.utfpr.edu.br

Resumo

Para conter a proliferação da pandemia de Covid-19, governos decretaram o fechamento de escolas, parques, igrejas, comércios e todo evento ou local que pudesse gerar aglomeração. Em decorrência dessa mudança de rotina nas cidades, o comportamento dos crimes tende a alterar, e compreender essa mudança é fundamental. Este estudo analisa o impacto da pandemia na incidência de crimes em Curitiba, em específico violência, substâncias ilícitas e roubo. É utilizado um modelo de regressão linear, que é implementado para compreender esse impacto. Ocorrências relacionadas à violência aumentaram durante o período de isolamento. De maneira oposta, os registros de roubo e furto foram menores. Por fim, ocorrências referentes à substâncias ilícitas não tiveram impactos significativos.

Abstract

To contain the spread of the Covid-19 pandemic, governments decreed the closing of schools, parks, churches, businesses and any event or venue that could generate crowding. As a result of this change in routine in cities, the behavior of crimes tends to change, and understanding this change is fundamental. This study analyzes the impact of the pandemic on the incidence of crimes in Curitiba, specifically violence, illegal substances and theft. A regression model is used, which is implemented to understand this impact. Occurrences related to violence increased during the isolation period. In the other hand, the records of burglary and robbery were lower. Occurrences about illegal substances had no significant impact.

Palavras-chaves: Criminalidade, Curitiba, Pandemia, Regressão, Série temporal

1. Introdução

No dia 11 de março de 2020 o diretor geral da Organização Mundial da Saúde (OMS), Tedros Adhanom, declarava que a Covid-19, uma infecção respiratória aguda causada pelo Coronavirus SARS-CoV-2, descoberta em pacientes com pneumonia na cidade de Wuhan,

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

na China, passaria a ser uma pandemia [1]. O mundo todo teve sua rotina alterada inesperadamente devido ao rápido contágio do vírus. Medidas drásticas de isolamento tiveram que ser tomadas por parte do poder público e privado para que os níveis de contaminação e disseminação diminuíssem.

Em Curitiba, o primeiro decreto para auxiliar na contenção da manifestação do vírus foi estabelecido em 16 de março de 2020, resultando no fechamento de escolas, comércios, templos, igrejas, shopping centers e qualquer evento ou local que pudesse gerar aglomeração. O efeito disso foi rapidamente sentido nos mais variados setores, inclusive na segurança pública. Com o isolamento estabelecido na cidade, é de se imaginar que o comportamento e distribuição dos mais variados crimes também fossem afetados. Segundo Cohen e Felson [2], a maioria dos crimes só ocorre quando indivíduos motivados a cometer alguma ilegalidade encontram alvos adequados e sem a presença de tutores, que poderiam interferir ou interromper o crime.

Tentar identificar algum padrão nessa interação entre período pandêmico, isolamento social e a variação nos números da criminalidade, pode auxiliar agentes públicos na suas estratégias e abordagens. Políticas públicas podem ser criadas com o objetivo de orientar e conscientizar a população. Ao identificar-se um aumento ou diminuição de crimes em uma determinada parte da cidade, políticas de seguranças podem ser aplicadas especificamente naquele local. Ações combativas podem ser direcionadas por bairro ou regional.

O objetivo deste trabalho é analisar e compreender o impacto da pandemia do Coronavirus e do isolamento social sobre a criminalidade em Curitiba, com base nas ocorrências registradas pela Guarda Municipal, com ênfase nos crimes de roubo, furto, violência e substância ilícita. São utilizados regressão linear e técnicas de análises de séries temporais.

2. Trabalhos relacionados

Desde o início da pandemia do Coronavírus, diversos estudos vêm sendo realizados para compreender o fenômeno da criminalidade em período de isolamento social e restrições rígidas de circulação. Mohler et al. [3] analisam o impacto do isolamento social durante a pandemia nos crimes em Los Angeles e Indianápolis, Estados Unidos. Primeiramente, os autores utilizaram uma regressão para cada tipo de incidente e, pelo fato do isolamento social ocorrer em diferentes datas e níveis para cada cidade, foi feita uma segunda regressão utilizando os índices de mobilidade do Google. De modo geral, o isolamento social pouco impactou nos números registrados em ambas as cidades. O número de chamados realizados referente à violência doméstica teve um aumento significativo tanto em Los Angeles quanto em Indianapolis. Por outro lado, roubo e furto tiveram um declínio modesto no número de chamados durante o período de isolamento.

Na Índia, um estudo realizado por Poblete-Cazenave [4] buscou analisar o efeito do isolamento social na incidência de crimes. Para isso, o autor implementou um modelo de regressão descontínua a fim de identificar alguma mudança em função do isolamento. O governo indiano classificou cada distrito em três categorias diferentes: verde, laranja e ver-

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

melho, sendo a última a mais restritiva. O estudo identificou que quanto mais severo é o nível de restrição, mais se intensifica a violência contra mulheres. Crimes violentos não tiveram mudanças drásticas, já crimes motivados economicamente aumentaram, sugerindo que o impacto econômico produzido pelo isolamento pode ter causado efeitos.

3. Metodologia

O estudo utiliza os dados disponibilizados pela Prefeitura Municipal de Curitiba, armazenados no Portal de Dados Abertos. Essa base contém informações das ocorrências registradas no Sistema de Gestão da Guarda Municipal, SiGesGuarda. Uma segunda base de dados foi utilizada para coletar informações acerca da quantidade populacional e rendimento mensal médio por responsáveis dos domicílios de cada bairro, permitindo, assim, uma análise do impacto da pandemia na criminalidade sob diferentes aspectos econômicos e populacionais.

Os crimes foram agrupados em três conjuntos diferentes, sendo eles

1. Roubo, onde foram agrupados as ocorrências de "Roubo" e "Furto";
2. Violência, que contém os registros de "Violação de Medida Protetiva Lei Maria da Penha","Ameaça", "Importunação Sexual", "Estupro", "Exploração de menores", "Maus tratos à pessoas" e "Abuso de incapazes";
3. Substância ilícita.

Dessa forma, as médias mensais de ocorrências registradas pela Guarda Municipal, desde 2018 até junho de 2021, foram de 178, 105 e 136 para roubo, violência e substância ilícita, respectivamente. Além disso, foi utilizada a biblioteca *statsmodels*, um módulo *Python* que fornece ferramentas para análises estatísticos.

Foi implementada, seguindo o modelo realizado em Los Angeles e Indianápolis, proposto por Mohler et al. [3], uma regressão em cada conjunto de crimes citados anteriormente. Nessa regressão, foram utilizadas uma variável binária, que, à partir de 16 de março de 2020, recebe o valor 1, e uma segunda variável categórica para o dia da semana, para controle de efeitos sazonais. Abaixo, a regressão executada para cada grupo de crime i:

$$Y_i = \beta_0 + \beta_1 \times isolamento + \beta_2 \times diaSemana \tag{1}$$

4. Resultados

Analisamos o impacto do isolamento social nos volumes de ocorrências de crimes ocorridos entre o dia 02 de janeiro de 2020 até 21 de abril de 2020, período idêntico ao utilizado em uma das regressões aplicadas no artigo Mohler et al. [3], a fim de comparar os resultados entre Curitiba, Los Angeles e Indianápolis, na tabela 1. Nas três cidades, a data de 16 de março de 2020 foi utilizada para demarcar o início do período de isolamento. Para controlar os efeitos sazonais na regressão, foi utilizada uma variável categórica, indicando o dia da semana. Além disso, uma segunda variável binária indica se havia ou não isolamento naquela data.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Tabela 1: Regressão de chamadas diárias para cada grupo de crime que tem como variável independente o dia da semana e se há isolamento.

Crime	Intercepto	Coef. Isolamento	Erro Padrão	P-Value
Roubo	7.9772	**-1.6059**	0.6108	**0,0098***
Violência	1.3919	**0.7881**	0.2944	**0,0086***
Substância	4.724	0.5975	0.524	0.257

Resultados significativos em negrito.
* Significativo após correção de Bonferroni.

Nos registros do grupo roubo, é possível notar que, a partir do momento em que foi estabelecido o fechamento geral de comércios e estabelecimentos públicos e privados, esse número foi estatisticamente menor que no período pré-fechamento, seguindo a tendência identificada em outros cidades e países, conforme apresentado anteriormente, inclusive Los Angeles e Indianápolis, mesmo utilizando o método de correção de Bonferroni, para 3 modelos diferentes, corrigindo o valor crítico de 0,05 para 0,016.

O impacto do isolamento social também refletiu nos números de ocorrências registradas relacionado ao grupo violência, dessa vez, em direção oposta. O número de chamados foi estatisticamente maior no período de isolamento. Essa diferença de média vai de encontro com a hipótese levantada no início deste trabalho, visto que a permanência em locais compartilhados entre agressor e vítima aumentou durante a pandemia.

Quando analisamos o impacto do isolamento social nos números de ocorrências de substâncias ilícitas, não encontramos, estatisticamente, alguma alteração significativa na diferença de médias nos dois períodos analisados.

Outra abordagem utilizada para análise foi a comparação entre as dez regionais de Curitiba, a fim de identificar certos comportamentos em regiões com população e renda distintas. A figura 1 está agrupada por mês, de janeiro de 2020 até julho de 2021. Além disso, as ocorrências foram normalizadas, dividindo o total de ocorrências de cada mês pela população da regional e multiplicando por cem mil.

Na figura 1, sobre roubo, é possível identificar uma queda acentuada na regional Matriz, região central de Curitiba, área com mais estabelecimentos comerciais e com um grande volume de pessoas circulando em dias normais, a partir de março de 2020, mês que foi decretado situação de emergência na cidade. Essa queda se repete, não com tanta intensidade, nas demais regionais. Com o passar dos meses, é possível notar, principalmente na regional Matriz, um aumento gradativo nos casos, o que mostra que esse crime se adaptou com a rotina que foi, aos poucos, voltando ao normal.

Ao analisar a figura 2, sobre violência, passados alguns meses após o decreto do isolamento e distanciamento social, notam-se alguns picos de ocorrências em algumas regionais. Tatuquara, que tem a menor renda média por responsáveis dos domicílios dentre todas as regionais, e Bairro Novo, com a terceira menor renda média, demonstram números altos no segundo semestre de 2020.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Figura 1: Análise por Regional - Roubo.

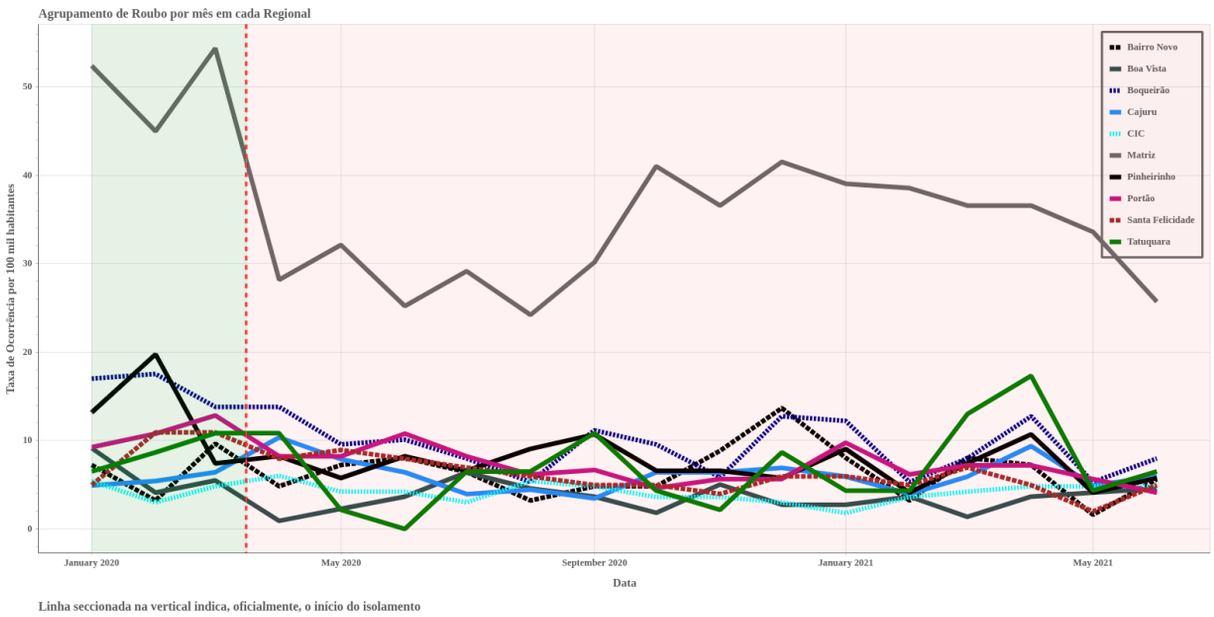
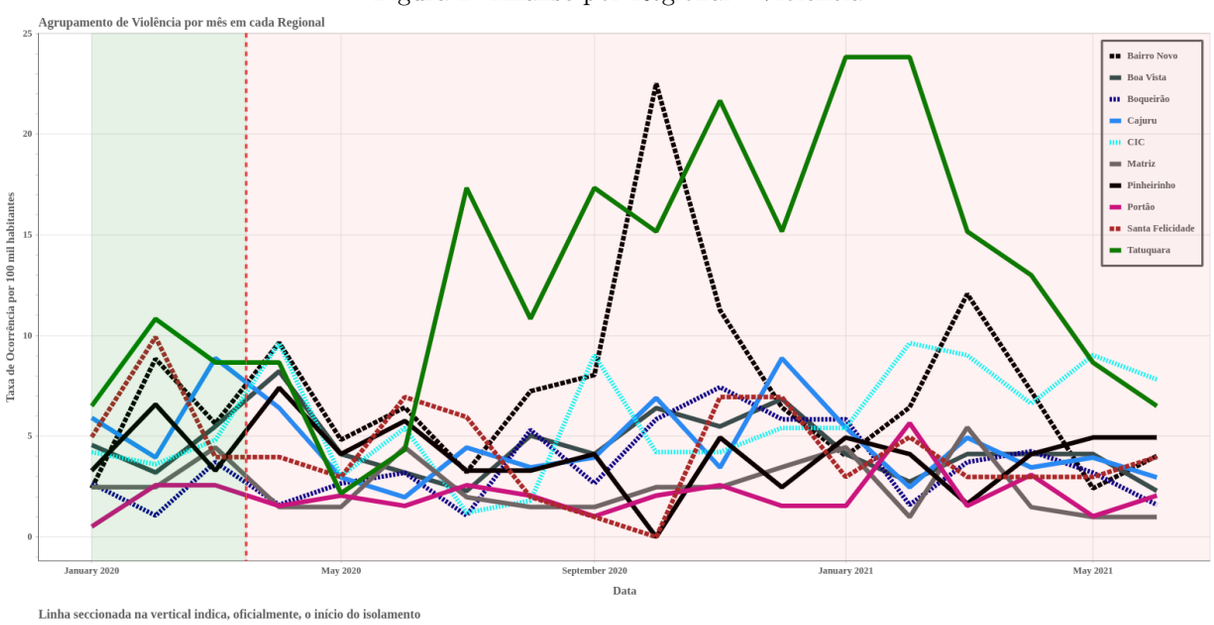


Figura 2: Análise por Regional - Violência.



5. Considerações Finais

Este trabalho apresentou uma análise do comportamento de três conjuntos de crimes antes e durante o período de pandemia: roubo, que também engloba furto, violência, que

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

agrupou crimes como violência doméstica, importunação sexual, violação de medida protetiva lei Maria da Penha, agressão física e verbal, entre outros, e substância ilícita.

Os resultados obtidos mostram que os crimes do conjunto violência aumentaram durante a pandemia. O fato do agressor e vítima estarem mais tempo juntos pode estar relacionado com esse aumento. Um ponto interessante identificado, e que pode ser melhor explorado em estudos futuros, é que as regionais Bairro Alto e Tatuquara, que estão entre as três regionais com a menor renda média, aumentaram suas taxas de ocorrências relacionadas a "violação de Medida Protetiva Lei Maria da Penha" em 60.5% e 20.8%, respectivamente, alguns meses após o período de isolamento. Matriz e Portão, que têm os dois melhores indicadores de renda média, diminuíram em 5.5% e 15.5%, respectivamente.

Roubo, conforme hipótese levantada no início do estudo, diminuiu. Essa queda foi mais acentuada na região central, local que havia um grande movimento comercial e circulação de pessoas antes da pandemia. Por fim, ocorrências relacionadas à substância ilícita não tiveram impacto considerável. Essa conclusão pode sugerir que, mesmo com o fechamento de fronteiras, o tráfico de drogas e o crime organizado continuaram a operar como antes da pandemia, e o isolamento não foi um empecilho nestes casos.

Trabalhos futuros podem explorar os dados da Guarda Municipal em conjunto com os da Polícia Civil e Militar, visto que crimes mais graves geralmente são denunciados através do telefone 190. Com esse agrupamento, resultados mais consistentes podem ser identificados.

Referências

[1] H. S. d. Farias, O avanço da covid-19 e o isolamento social como estratégia para redução da vulnerabilidade, Espaço e Economia. Revista brasileira de geografia econômica (2020).

[2] L. E. Cohen, M. Felson, Social change and crime rate trends: A routine activity approach, American Sociological Review 44 (1979) 588–608.

[3] G. Mohler, A. L. Bertozzi, J. Carter, M. B. Short, et al., Impact of social distancing during covid-19 pandemic on crime in los angeles and indianapolis, Journal of Criminal Justice 68 (2020) 101692.

[4] R. Poblete-Cazenave, The impact of lockdowns on crime and violence against women–evidence from india (2020).

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

An Extended Adjacency Lemma for Graph Edge-Coloring

Thiago H. F. M. Cunha[a,b], Leandro M. Zatesko[a,b]

[a] *Academic Department of Informatics, Federal University of Technology — Paraná, Curitiba, Brazil*
[b] *Emails: tcunha@alunos.utfpr.edu.br, zatesko@utfpr.edu.br*

Abstract

The study of edge-coloring emerged from the context of the well-known Four Color Problem. By Vizing's theorem, the minimum number of colors needed to color the edges of a simple graph G is either its maximum degree Δ, or it is $\Delta + 1$, in which case G is called Class 1, or Class 2. A critical graph is a connected Class 2 graph that becomes Class 1 by the removal of any edge. Vizing's recoloring procedure used in the proof of his theorem also yields a condition, known as Vizing's Adjacency Lemma, for the minimum number of vertices of degree Δ adjacent to any vertex in a critical graph. We demonstrate an extension of this adjacency lemma using an extended recoloring procedure presented recently.

Keywords: Coloring of graphs and hypergraphs, Edge subsets with special properties, Graph algorithms
2020 MSC: 05C15, 05C70, 05C85

1. Introduction

Throughout this text, all graphs are simple. Let G be a graph, wherein $V(G)$ is the vertex set and $E(G)$ is the edge set of G. An edge uv in $E(G)$ is formed by two endpoints, $u, v \in V(G)$. The *degree* $d_G(u)$ of $u \in V(G)$ is the number of edges incident with u. The set of neighbors of u in G is denoted $N_G(u)$. Two edges are adjacent if they both are incident with some $u \in V(G)$. The maximum (minimum) degree of G is denoted $\Delta(G)$ ($\delta(G)$). Whenever free of ambiguity, we can omit the graph G from the notation, writing simply $d(u)$, Δ etc.

Vertices of maximum degree in G are referred to as the *majors* of G. The *core* of G, denoted $\Lambda[G]$, is the subgraph of G induced by its majors. The majors u of G with $\sum_{v \in N_G(u)} d_G(v) \geq \Delta^2 - \Delta + 2$ are the *proper* majors of G. If this sum is equal to $\Delta^2 - \Delta + 1$, then u is a *tightly non-proper* major of G. Otherwise, u is a *strictly non-proper* major of G. The *hard core* of G, denoted $\mathbf{\Lambda}[G]$, is the subgraph of G that is induced by all proper and tightly non-proper majors of G. Also, the *soft core* of graph G is the subgraph of G induced by the strictly non-proper majors of G.

An *edge-coloring* of a graph G is an assignment of colors to the edges of G so that no two adjacent edges are colored the same. For two colors α, β used in an edge-coloring, the *α/β-component* is the subgraph of G induced by the edges colored α or β. We say that a vertex u misses some color α if no edge incident with u in colored α.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

The *chromatic index* of a graph G, denoted $\chi'(G)$, is the least amount of colors needed for an edge-coloring of G. The Edge-coloring Problem consists of finding the chromatic index of a given graph. This problem emerges from the context of the well-known Four Color Problem, which is equivalent to the statement that every planar cubic graph has chromatic index equal to three [1]. By Vizing's theorem [2], the chromatic index of a simple graph G is either $\Delta(G)$ or $\Delta(G)+1$, being G called Class 1 or Class 2, respectively. Deciding if a graph G is Class 1 is NP-complete [3].

The proof of Vizing's theorem follows from the recoloring procedure established in Definition 1 and Lemma 1. In both statements, G is a graph for which we already have an edge-coloring of $G - uv$ using t colors, for some $uv \in E(G)$ and some $t \in \mathbb{Z}_{>0}$.

Definition 1. A *Vizing's recoloring fan* for uv is a sequence $v_0, v_1, \ldots, v_k$ of distinct neighbors of u in G such that $v_0 = v$ and, for all $i \in \{0, \ldots, k-1\}$, vertex v_i misses the color α_i, which is the color of edge uv_{i+1}. This fan is said to be *complete* if both v_k and u miss the same color.

Lemma 1 ([2])**.** *If there is a Vizing's recoloring fan $v_0, v_1, \ldots, v_k$ for uv satisfying at least one of the properties below, then G also admits an edge-coloring using t colors.*

Properties 1a. *The fan $v_0, v_1, \ldots, v_k$ is complete.*

Properties 1b. *Vertex v_k misses the same color a_j as v_j in G for some $j < k$.*

With Vizing's recoloring procedure, we have another important result, known as Vizing's Adjacency Lemma. In the statement, a *critical* graph is a connected Class 2 graph G such that $\chi'(G - uv) < \chi'(G)$ for all $uv \in E(G)$.

Lemma 2 (Vizing's Adjacency Lemma [4])**.** *For every edge uv of a critical graph G, the number of majors adjacent to u in G is at least:*

$$\begin{cases} \Delta - d_G(v) + 1, & \text{if } d_G(v) < \Delta; \\ 2, & \text{if } d_G(v) = \Delta. \end{cases}$$

Recently, an extension of Vizing's recoloring procedure was proposed [5]. Definition 2 and Lemma 3 describe this procedure. In both statements, G is a graph for which we already have an edge-coloring of $G - uv$ using Δ colors, for some $uv \in E(G)$.

Definition 2. An *extended recoloring fan* for uv is a sequence $v_0, v_1, \ldots, v_k$ of distinct neighbors of u in G such that $v_0 = v$ and, for all $i \in \{0, \ldots, k-1\}$, either v_i actually misses color α_i, or $i > 0$ and v_i misses this color *virtually*, that is, there is an edge $v_i w_i$ in G colored α_i such that $w_i \neq v_{i-1}$ actually misses α_{i-1}. This fan is said to be *complete* if v_k (actually or virtually) misses some color that is also missing at u (see Figure 1).

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

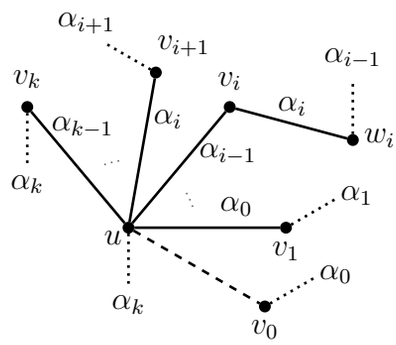


Figure 1: A complete extended recoloring fan

Lemma 3 ([5])**.** *If there is an extended recoloring fan $v_0, v_1, \ldots, v_k$ for uv satisfying at least one the properties below, then G is Class* 1*.*

Properties 3a. *The fan $v_0, v_1, \ldots, v_k$ is complete.*

Properties 3b. *Vertex v_k misses, virtually or actually, the same color a_j as v_j in G for some $j < k$.*

If there is an extended recoloring fan $v_0, v_1, \ldots, v_k$ that satisfies Property 3a, then the edge-coloring of G with Δ colors can be obtained by a procedure referred to as *decay of the colors* of the fan, which is performed for i from k down to 0. At the beginning of each iteration, it is invariant that both u and v_i miss α_i (the latter possibly virtually). So, uv_i is colored α_i and, if v_i misses α_i virtually, also v_iw_i is colored α_{i-1}. If $i = 0$, we are done. If $i > 0$, now u misses α_{i-1}, which is still missing (possibly virtually) at v_{i-1}, so we can decrement i and continue.

Lemma 4 brings another interesting application of the extended recoloring procedure.

Lemma 4 ([5])**.** *Let G be a graph for which we already have an edge-coloring of $G - uv$ using Δ colors, for some $uv \in E(G)$. If all majors adjacent to u in $G - uv$ are strictly non-proper, then G is Class* 1*.*

In this paper, we show a stronger form of Lemma 2 with the use of the extended recoloring procedure.

2. Result

Lemma 5 (Extended Adjacency Lemma)**.** *For every edge uv of a critical graph G, the number of vertices in $\mathbf{\Lambda}[G]$ adjacent to u is at least:*

$$\begin{cases} \Delta - d_G(v) + 1, & \text{if } d_G(v) < \Delta; \\ 2, & \text{if } d_G(v) = \Delta. \end{cases}$$

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Proof. Consider the following:

Claim. Let uv be an edge of a graph G such that $\chi'(G-uv) = \Delta(G-uv) = \Delta(G) =: \Delta$. If u is adjacent in $G-uv$ to at most $\Delta - d_G(v)$ vertices in $\mathbf{\Lambda}[G]$, then G is Class 1.

First, we will show that the lemma follows directly from the claim. Let G be a critical simple graph and uv be any edge of G. By the definition of critical graphs, it is known that $\chi'(G-uv) \leq \Delta$. We also know that $\Delta(G-uv) = \Delta(G)$, since otherwise every vertex of $G-uv$ would miss a color and, thus, we would be able to construct an edge-coloring for G using $\Delta(G)$ colors by Vizing's recoloring procedure. For the sake of contradiction, suppose that the number of vertices in $\mathbf{\Lambda}[G]$ adjacent to u in G is at most:

$$\begin{aligned} \Delta - d_G(v), &\quad \text{if } d_G(v) < \Delta; \\ 1, &\quad \text{if } d_G(v) = \Delta. \end{aligned}$$

Note that, in the case wherein $d_G(v) = \Delta$, if $v \in V(\mathbf{\Lambda}[G])$ then v is the only major in $\mathbf{\Lambda}[G]$ adjacent to u in G. Also note that, in both cases, we have that u is adjacent in $G-uv$ to at most $\Delta - d_G(v)$ vertices in $\mathbf{\Lambda}[G]$. Ergo, by the claim, G is Class 1, but this is a contradiction, because G is critical graph, thus Class 2 by definition.

Now, we will prove the claim. For this, consider an edge-coloring of $G-uv$ using Δ colors.

If $d_G(v) = \Delta$, then u is adjacent in $G-uv$ to no majors in $\mathbf{\Lambda}[G]$. Therefore, all majors adjacent to u in $G-uv$ are strictly non-proper and edge uv is not colored, so we can apply Lemma 4 to obtain an edge-coloring of G using Δ colors.

If $d_G(v) < \Delta$, we know that the numbers of colors missing at u is at least one, and at $v := v_0$ is exactly $\Delta - d_G(v) + 1 \geq 2$. If there is some extended recoloring fan $v_0, v_1, \ldots, v_k$ satisfying any of Properties 3a, 3b, we are done. Otherwise, let $v_0, v_1, \ldots, v_k$ be a maximal recoloring fan for uv. Clearly, v_k is major of G. If $v_k \notin V(\mathbf{\Lambda}[G])$, then it is always possible to recolor some edges of G either to obtain a color virtually missing at v_k, or a complete recoloring fan for uv [5]. Therefore, we assume without loss of generality that for every maximal recoloring fan $v_0, v_1, \ldots, v_k$ for uv, vertex $v_k \in V(\mathbf{\Lambda}[G])$.

For each of the $\Delta - d_G(v) + 1$ missing colors at v_0, we have a distinct maximal recoloring fan starting at v_0. Since u is adjacent to at most $\Delta - d_G(v)$ vertices in $\mathbf{\Lambda}[G]$, we have, by the Pigeonhole Principle, two distinct maximal recoloring fans $v_0, x_1, , \ldots, x_m$ and $v_0, y_1, \ldots, y_n$ such that $x_i = y_j$ for some $i \in \{2, \ldots, m\}$ and some $j \in \{2, \ldots, n\}$. Without loss of generality, we can assume that $x_{i'} \neq y_{j'}$ for all $i' \in \{1, \ldots, i-1\}$ and for all $j' \in \{1, \ldots, j-1\}$.

Let α be the color of $ux_i = uy_j$ and let β be some color missing at u. Initially, uncolor edge ux_i. Now u misses α and, since both x_{i-1} and y_{j-1} miss α, we have two distinct complete fans $v_0, x_1, \ldots, x_{i-1}$ and $v_0, y_1, \ldots, y_{j-1}$. Choose one of them to perform the decay of the colors, say $v_0, x_1, \ldots, x_{i-1}$.

We know that v_0 misses at least two colors, thus, after the decay, we can rearrange the fans to obtain the fan

$$x_i, x_{i-1}, \ldots, x_2, x_1, v_0, y_1, y_2, \ldots, y_{j-2}, y_{j-1}$$

which is a recoloring fan for ux_i satisfying Property 3b, since both x_i and y_{j-1} miss α. □

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

3. Final remarks

Many results in the literature concerning edge-coloring are derived from Vizing's Adjacency Lemma [4, 6]. With the use of the recent extended recoloring procedure presented in [5], we sought to prove a stronger form of Vizing's Adjacency Lemma. Similarly, we aim to investigate if, through the extended recoloring procedure and the extended adjacency lemma proved in this work, we can also achieve stronger forms of other known results.

Acknowledgements

We thank the anonymous referee for the valuable suggestions given.

References

[1] P. G. Tait, On the colouring of maps, Proc. Roy. Soc. Edinburgh Sect. A 10 (1878–1880) 501–503,729.

[2] V. G. Vizing, On an estimate of the chromatic class of a p-graph (in Russian), Diskret. Analiz. 3 (1964) 25–30.

[3] I. Holyer, The NP-completeness of edge-colouring, SIAM J. Comput. 10 (1981) 718–720.

[4] V. G. Vizing, Critical graphs with a given chromatic class (in Russian), Diskret. Analiz. 5 (1965) 9–17.

[5] L. M. Zatesko, A. Zorzi, R. Carmo, A. L. P. Guedes, Edge-colouring graphs with bounded local degree sums, Discrete Appl. Math. 281 (2020) 268–283. URL: `https://doi.org/10.1016/j.dam.2020.03.044`.

[6] J.-C. Fournier, Méthode et théorème générale de coloration des arêtes, J. Math. Pures Appl. 56 (1977) 437–453.

Proper $\mathscr{L}(h,k)$-labelling of Caterpillars and Multisunlets[1]

J. P K. Castilho[a,c], C. N. Campos[a], L. M. Zatesko[b]

[a] *IC, University of Campinas, Brazil*
[b] *DAINF, Federal University of Technology – Paraná, Brazil*
[c] *Corresponding author: joao.pkc@gmail.com*

Abstract

An $\mathscr{L}(h,k)$-*labelling* of a simple graph G is a function $\sigma\colon V(G) \to \mathbb{Z}_{\geq 0}$ such that the labels of: adjacent vertices are at least h apart; vertices which have a common neighbour are at least k apart. The *span* of σ is the largest difference between the labels of any two vertices. Computing the $\mathscr{L}(h,k)$-span of G, the least span amongst all σ, is NP-hard even for trees. For $h \geq k$, we determine the $\mathscr{L}(h,k)$-span of the *Uniform Multisunlets* (*Uniform Caterpillars*), in which $p \geq 1$ pendant vertices are added at each vertex of a basis cycle (path).

Keywords: Graph labelling, Caterpillar graphs, Sunlet graphs

1. Introduction

Let G be a simple graph with *vertex* set $V(G)$ and *edge* set $E(G)$. An edge $e \in E(G)$ with *endpoints* u and v is denoted uv. The *degree* of $u \in V(G)$ is denoted $d(u)$. A *pendant* is a vertex of degree one. The set of *neighbours* of u is denoted $N(u)$.

The $\mathscr{L}(2,1)$-labelling, proposed to approach the Frequency Assignment Problem [1], is an assignment of integers to the vertices of a graph wherein adjacent vertices receive labels that are at least two apart and non-adjacent vertices having a common neighbour receive different labels. Later, Georges and Mauro [2] generalised this concept as presented in Definition 1.

Definition 1. Let $h, k \in \mathbb{Z}_{\geq 0}$ and G be a simple graph. An $\mathscr{L}(h,k)$-*labelling* of G is a function $\sigma\colon V(G) \to \mathbb{Z}_{\geq 0}$ such that:

(i) $|\sigma(u) - \sigma(v)| \geq h$, for all $uv \in E(G)$;

(ii) $|\sigma(u) - \sigma(v)| \geq k$, for all distinct $u, v \in V(G)$ with $N(u) \cap N(v) \neq \emptyset$.

The *span* of an $\mathscr{L}(h,k)$-labelling σ is $\lambda(\sigma) = \max_{u,v \in V(G)}\{\sigma(u) - \sigma(v)\}$. The $\mathscr{L}(h,k)$-*span* of G is $\lambda_{h,k}(G) = \min_{\sigma}\{\lambda(\sigma)\}$. Some authors define (ii) in Definition 1 as $|\sigma(u) - \sigma(v)| \geq k$ *for all* $u, v \in V(G)$ *with* $dist(u,v) = 2$. Note that the two definitions are equivalent when

[1] Partially supported by CNPq (425340/2016-3 and 428941/2016-8) and CAPES.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

$h \geq k$ or when G is triangle-free. On that account, we say that σ is a *proper* $\mathscr{L}(h,k)$-labelling if $h \geq k$.

The problem of determining the exact value of the $\mathscr{L}(h,k)$-span is completely settled only for basic graph classes, such as stars, paths, and cycles, or specific values of h and k [2, 3]. On the other hand, this problem is NP-hard even for trees whenever k does not divide h [4].

Besides paths and stars, another important subclass of trees are the *Caterpillars*. These are trees for which there is a path P such that every vertex either lies in P or is adjacent to a vertex in P. We approach the proper $\mathscr{L}(h,k)$-labelling problem in this class, determining the $\mathscr{L}(h,k)$-span of *Uniform Caterpillars* $\overset{\vee}{P}_{n,p}$, in which $p \geq 1$ pendant vertices are added at each vertex of a path of n vertices, called the *basis* of $\overset{\vee}{P}_{n,p}$, as in Figure 1.

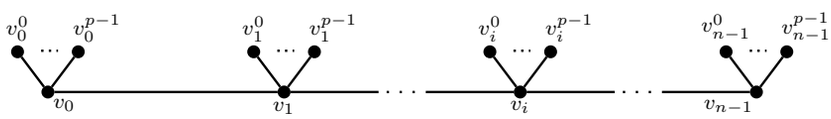


Figure 1: The Uniform Caterpillar $\overset{\vee}{P}_{n,p}$.

As Caterpillars generalise paths, an interesting generalisation of the cycles are the *Sunlets* $\overset{\prime}{C}_n$, also much studied in labelling problems [5]. These are the graphs obtained by adding one pendant vertex at each vertex of a cycle of n vertices, called the *basis* of $\overset{\prime}{C}_n$. We not only settle the proper $\mathscr{L}(h,k)$-labelling problem for the Sunlets, but we also determine the span of *Uniform Multisunlets* $\overset{\vee}{C}_{n,p}$, in which $p \geq 1$ pendant vertices are added at each vertex of a cycle of n vertices, the *basis* of $\overset{\vee}{C}_{n,p}$.

2. Preliminaries

In an $\mathscr{L}(h,k)$-labelling, the set of all neighbours of some $u \in V(G)$ whose labels are smaller (greater) than the label of u is denoted $N_<(u)$ ($N_>(u)$). For $n,p \in \mathbb{Z}_{\geq 0}$, the vertices of the basis of $\overset{\vee}{P}_{n,p}$ ($\overset{\vee}{C}_{n,p}$) are denoted v_i for $0 \leq i < n$, and the pendant vertices at each v_i are denoted v_i^j for $0 \leq j < p$, as in Figure 1. For $n \in \mathbb{Z}_{\geq 2}$, the *centre* of the *star* $K_{1,n}$ is the vertex with degree n. We use $[a \,..\, b]$ to denote the set $\{i \in \mathbb{Z} : a \leq i \leq b\}$.

Since an $\mathscr{L}(h,k)$-labelling of a graph G induces an $\mathscr{L}(h,k)$-labelling of every subgraph of G, and since each vertex of the basis of $\overset{\vee}{P}_{n,p}$ ($\overset{\vee}{C}_{n,p}$) and its neighbours induce a subgraph with a spanning star, we establish some properties of $\mathscr{L}(h,k)$-labellings of stars with special values of span.

Theorem 1 ([3]). *Let $h,k \in \mathbb{Z}_{\geq 0}$ and $p \in \mathbb{Z}_{>0}$ such that $h \geq k$. Then, $\lambda_{h,k}(K_{1,p}) = h + (p-1)k$.* □

Lemma 1. *Let $\alpha,\beta,i,h,k,p \in \mathbb{Z}_{\geq 0}$ with: either $\alpha = 1$, $i \geq 2$, $h \geq ik$, and $\beta \leq p+i$; or $\alpha = 2$ and $\beta \leq p$. If there is an $\mathscr{L}(h,k)$-labelling σ of $K_{1,p+2}$ with span $\alpha h + \beta k - 1$, then, being u the centre of $K_{1,p+2}$, one of the following holds: $N_<(u) = \emptyset$ and $\sigma(u) < (\alpha-1)h + (\beta-p-1)k$; $N_>(u) = \emptyset$ and $\sigma(u) \geq h + (p+1)k$.* □

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Lemma 2. *Let $h, k, a, \gamma, p \in \mathbb{Z}_{\geq 0}$ with $0 < k < h < 2k$, $a = \max\{0, \lceil 3k/2 \rceil - h\}$, and: $\gamma = 1$ if $3k$ is odd and $h < 3k/2$; $\gamma = 0$ if otherwise. Let $\mathcal{X}_i = [ik - 2a + \gamma \,..\, h + (i-1)k - 1]$ for $i \in [1 \,..\, p+2]$. If there is an $\mathscr{L}(h,k)$-labelling of $K_{1,p+2}$ with span $\min\{h + (p+2)k - 1, 3h + (p-1)k - 1\}$, then the label of the centre of $K_{1,p+2}$ does not belong to $\mathcal{X}_i$ for any i.* □

Lemma 3. *Let $h, k, a, \gamma, p \in \mathbb{Z}_{\geq 0}$ with $0 < k < h < 2k$, $a = \max(0, \lceil 3k/2 \rceil - h)$, and: $\gamma = 1$ if $3k$ is odd and $h < 3k/2$; $\gamma = 0$ if otherwise. Let σ be an $\mathscr{L}(h,k)$-labelling of $K_{1,p+2}$ such that $\lambda(\sigma) = \min\{h + (p+2)k - 1, 3h + (p-1)k - 1\}$. Let $\mathcal{Y}_i = [h + (i-1)k \,..\, (i+1)k - 2a + \gamma - 1]$ for $i \in [1 \,..\, p+1]$. Let u be the centre of $K_{1,p+2}$. If $\sigma(u) \in \mathcal{Y}_i$ for some $i \in [1 \,..\, p+1]$, then, for all $v \in V(K_{1,p+2}) \setminus \{u\}$, $\sigma(v) \notin \mathcal{Y}_j$ for any $j \in [1 \,..\, p+1]$.* □

3. Main results

In Theorems 2 and 3, we establish $\lambda_{h,k}(\overset{\vee}{P}_{n,p})$ and $\lambda_{h,k}(\overset{\vee}{C}_{n,p})$, respectively.

Theorem 2. *Let $h, k, n, p \in \mathbb{Z}_{\geq 0}$ such that $h \geq k$, $p \geq 1$, and $n \geq 2$. Then, $\lambda_{h,k}(\overset{\vee}{P}_{n,p}) = \min\{h + (p+i)k, 2h + pk\}$, wherein $i = \min\{2, \lceil n/2 \rceil - 1\}$.*

Proof (sketch). If $2 \leq n \leq 4$, then $h + (p+i)k \leq 2h + pk$. In this case, the lower bound follows from Theorem 1, since $K_{1,p+i+1} \subseteq \overset{\vee}{P}_{n,p}$, and the upper bound follows by inspection on the labellings presented in Figure 2.

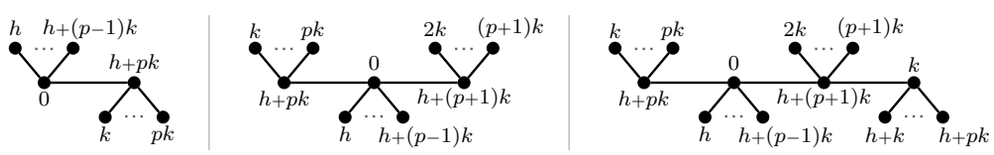


Figure 2: $\mathscr{L}(h,k)$-labelling of $\overset{\vee}{P}_{n,p}$ in the case wherein $2 \leq n \leq 4$

In order to show the lower bound for $n \geq 5$, assume that there is an $\mathscr{L}(h,k)$-labelling σ for $\overset{\vee}{P}_{n,p}$ with $\lambda(\sigma) = \min\{h + (p+2)k, 2h + pk\} - 1$. If $h + (p+2)k \leq 2h + pk$, then divide the available labels into three sets $\mathcal{X}_1 = [0 \,..\, k-1]$, $\mathcal{X}_2 = [k \,..\, h + (p+1)k - 1]$, and $\mathcal{X}_3 = [h + (p+1)k \,..\, h + (p+2)k - 1]$. Since $\sigma(v_0)$, $\sigma(v_1)$, and $\sigma(v_3)$ are centres of a $K_{1,p+2}$, by Lemma 1, these vertices cannot be labelled with elements of $\mathcal{X}_2$. Also, we cannot have two of these three vertices simultaneously labelled with elements of either $\mathcal{X}_1$ or $\mathcal{X}_3$. Therefore, one of these vertices cannot be assigned any label. The reasoning for case $2h + pk \leq h + (p+2)k$ is similar.

For the upper bound, we present an $\mathscr{L}(h,k)$-labelling σ with same span as in the statement for: $\overset{\vee}{P}_{3t,p}$, with $t = \lceil n/3 \rceil$, if $\lambda(\sigma) = 2h + pk$; $\overset{\vee}{P}_{4t,p}$, with $t = \lceil n/4 \rceil$, if $\lambda(\sigma) = h + (p+2)k$. This labelling is constructed by replicating the blocks of Figure 3. Observing that $\overset{\vee}{P}_{n,p}$ is a subgraph of $\overset{\vee}{P}_{N,p}$ for any $N \geq n$, we conclude $\lambda_{h,k}(\overset{\vee}{P}_{n,p}) \leq \min\{h + (p+2)k, 2h + pk\}$. □

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

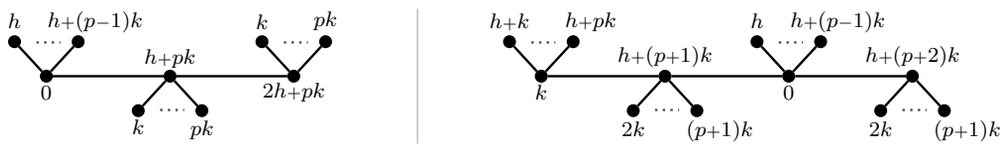


Figure 3: The blocks for our construction of an $\mathscr{L}(h,k)$-labelling with span: $2h+pk$ for $\overset{\vee}{P}_{3t,p}$ (left); $h+(p+2)k$ for $\overset{\vee}{P}_{4t,p}$ (right).

Theorem 3. *Let $h,k \in \mathbb{Z}_{\geq 0}$ such that:* (a) $0 \leq 2k \leq h$; *or* (b) $0 < k \leq h < 2k$. *Let $n,p \in \mathbb{Z}_{>0}$ such that $n \geq 3$. Let $i = (n \bmod 4)/2$. Then,*

$$\lambda_{h,k}(\overset{\vee}{C}_{n,p}) = \begin{cases} \min\{h+(p+2)k, 3h+(p-1)k+\lfloor 1/p \rfloor h\} & \text{if } n=5 \text{ and (b)}, \\ \min\{h+(p+2+i)k, 2h+pk\} & \text{if } n \text{ is even and (a)}, \\ 2h+pk & \text{in any other case.} \end{cases}$$

Proof (sketch). Initially, we consider $n \neq 5$. First, we construct an $\mathscr{L}(h,k)$-labelling with span $2h+pk$. For $n \in \{3,4,8\}$, the labelling is presented in Table 1. For $6 \leq n \leq 7$ and $n \geq 9$, we start from the labelling for: $\overset{\vee}{C}_{3,p}$ if $n \equiv 0 \pmod 3$; $\overset{\vee}{C}_{4,p}$ if $n \equiv 1 \pmod 3$; or $\overset{\vee}{C}_{8,p}$ if $n \equiv 2 \pmod 3$. Then, the remaining vertices are labelled in blocks of three, following Figure 4.

Table 1. An $\mathscr{L}(h,k)$-labelling of $\overset{\vee}{C}_{n,p}$ for $n \in \{3,4,8\}$. Here, $0 < j < p$.

$n \bmod 3$	v_0	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_0^0	v_0^j	v_1^0	v_1^j
0	0	$2h+pk$	$h+pk$	–	–	–	–	–	h	$h+jk$	k	$(j+1)k$
1	0	h	$2h+pk$	$h+pk$	–	–	–	–	$2h+pk$	$h+jk$	$2h$	$2h+jk$
2	0	h	$2h+pk$	$h+pk$	0	h	$2h+pk$	$h+pk$	$2h+pk$	$h+jk$	$2h$	$2h+jk$

$n \bmod 3$	v_2^0	v_2^j	v_3^0	v_3^j	v_4^0	v_4^j	v_5^0	v_5^j	v_6^0	v_6^j	v_7^0	v_7^j
0	k	$(j+1)k$	–	–	–	–	–	–	–	–	–	–
1	0	$h+jk$	k	$(j+1)k$	–	–	–	–	–	–	–	–
2	0	$h+jk$	k	$(j+1)k$	$2h+pk$	$h+jk$	$2h$	$2h+jk$	0	$h+jk$	k	$(j+1)k$

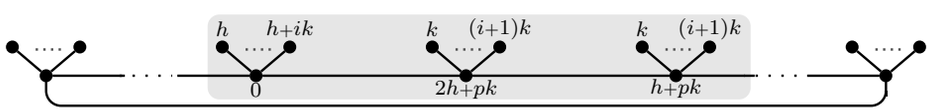


Figure 4: In grey, a building block for the labelling with span $2h+pk$ when $n \neq 5$.

For n even and $h \geq 2k$, we present another labelling, with span $h+(p+2+i)k$, wherein $i = (n \bmod 4)/2$. Note that $h+(p+2+i)k \leq 2h+pk$ if and only if $h \geq (2+i)k$. If $n \in \{4,6\}$, the labelling is presented in Table 2. If $n \geq 8$, we start from the labelling for: $\overset{\vee}{C}_{4,p}$ if $n \equiv 0$

Table 2. $\mathscr{L}(h,k)$-labelling of $\overset{\vee}{C}_{n,p}$ for $n \in \{4,6\}$. Here, $0 < j < p$.

n mod 4	v_0	v_1	v_2	v_3	v_4	v_5	v_0^0	v_0^j	v_1^0	v_1^j
0	0	$h+(p+2)k$	k	$h+(p+1)k$	–	–	h	$h+jk$	$2k$	$(j+2)k$
2	0	$h+(p+3)k$	k	$h+(p+1)k$	$2k$	$h+(p+2)k$	h	$h+jk$	$2k$	$(j+2)k$

n mod 4	v_2^0	v_2^j	v_3^0	v_3^j	v_4^0	v_4^j	v_5^0	v_5^j
0	$h+k$	$h+(j+1)k$	$2k$	$(j+2)k$	–	–	–	–
2	$h+k$	$h+(j+1)k$	0	$(j+2)k$	$h+(p+3)k$	$h+(j+1)k$	$3k$	$(j+3)k$

(mod 4); $\overset{\vee}{C}_{6,p}$ if $n \equiv 2 \pmod 4$. Then, for every $i \in [4+(n \bmod 4) \,..\, n-1]$, vertex v_i and its pendants are assigned the same labels of v_{i-4} and its pendants.

The proof of the lower bound is similar to that in Theorem 2. Being s the value established in the statement, we assume $\lambda(\overset{\vee}{C}_{n,p}) = s-1$ and partition the available labels into the three sets defined in Table 3, showing that at least one vertex of the basis cannot be assigned any label.

Table 3. Definition of sets $\mathcal{X}_1$, $\mathcal{X}_2$, and $\mathcal{X}_3$ when $n \neq 5$ or $h \geq 2k$ for cases: (i) n is even, $h \geq 2k$, and $i = (n \bmod 4)/2$; (ii) n in odd and $h \geq 2k$, or $n \neq 5$ and $h < 2k$.

	$\mathcal{X}_1$	$\mathcal{X}_2$	$\mathcal{X}_3$
(i)	$[0 \,..\, (1+i)k-1]$	$[(1+i)k \,..\, h+(p+1)k-1]$	$[h+(p+1)k \,..\, h+(p+2+i)k-1]$
(ii)	$[0 \,..\, h-k-1]$	$[h-k \,..\, h+(p+1)k-1]$	$[h+(p+1)k \,..\, 2h+pk-1]$

Consider, now, $n = 5$. The upper bound is established in Figure 5. The lower bound is quite challenging since the basis is a cycle with five vertices. By taking sets $\mathcal{X}_1, \ldots, \mathcal{X}_{p+2}$ and $\mathcal{Y}_1, \ldots, \mathcal{Y}_{p+1}$ as in Lemmas 2 and 3, we reach a contradiction using a similar technique to those presented so far. □

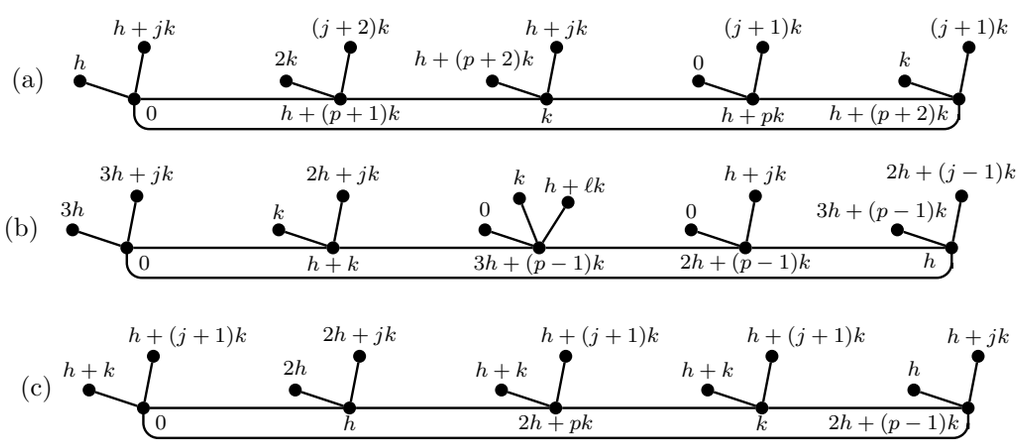


Figure 5: In (a)–(c), $0 < j < p$. In (a), $0 < k \leq h < 2k$ and: $p = 1$, or $p \geq 2$ and $h + (p+2)k \leq 3h + (p-1)k$; if $p = 1$, only the leftmost pendant exists. In (b), $0 < k \leq h < 2k$, $p \geq 2$, $3h + (p-1)k \leq h + (p+2)k$, and $1 < \ell < p$. In (c), $h \geq 2k$. □

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Acknowledgements

We thank the anonymous referee for the valuable suggestions given.

References

[1] J. R. Griggs, R. K. Yeh, Labelling graphs with a condition at distance 2, SIAM J. Discrete Math. 5 (1992) 586–595.

[2] J. Georges, D. Mauro, Generalized vertex labelings with a condition at distance two, Congr. Numer. 109 (1995) 141–159.

[3] T. Calamoneri, A. Pelc, R. Petreschi, Labeling trees with a condition at distance two, Discrete Math. 306 (2006) 1534–1539.

[4] J. Fiala, P. A. Golovach, J. Kratochvíl, Computational complexity of the distance constrained labeling problem for trees, in: Proc. ICALP '08, 2008, pp. 294–305.

[5] J. A. Gallian, A dynamic survey of graph labeling, Electron. J. Combin. 22 (2018) 1–535.

Grover's Quantum Algorithm and the Maximum Clique Problem

Matheus Giovanni Dias[a,b], Leandro Miranda Zatesko[a,b]

[a] *Academic Department of Informatics, Federal University of Technology — Paraná, Curitiba, Brazil*
[b] *Emails: matheusdias@alunos.utfpr.edu.br, zatesko@utfpr.edu.br*

Abstract

The Maximum Clique Problem consists of finding in a simple graph the largest subset of vertices wherein each pair of vertices is connected by an edge. It is an NP-Hard problem with many practical applications. In this paper we show how to implement an oracle for Grover's Quantum Search, yielding a quantum algorithm for finding the maximum clique in an n-vertex simple graph. The yielded algorithm has overall time complexity $\mathcal{O}(\sqrt{2^n}(n \log n)^2)$ and space complexity $\mathcal{O}(n^2)$.

Keywords: Vertex subsets with special properties, Quantum algorithms and complexity in the theory of computing, Graph algorithms
2020 MSC: 05C69, 68Q12, 05C85

1. Introduction

In graph theory, a *clique* in a simple graph is a subset of vertices wherein every pair of distinct vertices are adjacent (connected by an edge). Throughout this text, a graph is always a simple graph. A *maximal* clique is a clique that is not contained in a larger clique in the same graph. A *maximum* clique is a largest maximal clique of a graph. Determining the cardinality of the maximum clique in a graph is a well-known NP-Hard problem [1]. The best-known classical algorithm for this problem has time complexity $\mathcal{O}(3^{\frac{n}{3}})$. The Maximum Clique Problem is also hard to approximate [2], i.e. for any $\varepsilon > 0$, there cannot be a polynomial-time approximation algorithm for the problem within an $\mathcal{O}(n^{1-\varepsilon})$ approximation factor, unless $\mathsf{P} = \mathsf{NP}$.

Grover's Quantum Search is a quantum[1] search algorithm that provides a polinomial speedup over the best-known classical solutions for several important problems. This algorithm uses the quantum superposition and interference phenomena to search in a set of possible solutions for a value x that *satisfies* a black-box Boolean function $f : \{0,1\}^n \to \{0,1\}$, referred to as the *oracle* for the problem, that is, $f(x) = 1$ when x is a solution for the

[1] We refer the reader to: Nielsen and Chuang [3] for preliminaries on Quantum Computing and Quantum Mechanics; Diestel [4] for preliminaries on Graph Theory.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

problem, and $f(x) = 0$ otherwise. After $\mathcal{O}(\sqrt{2^n})$ evaluations of the function, the probability that a solution is found is $1 - \mathcal{O}(1/n)$.

Bojié [5] proposed an algorithm based on Grover's Quantum Search for finding a maximum clique in a simple graph G. He uses an iterative approach for searching a clique in a list of subgraphs with at least k vertices, starting at $k = 1$. For each iteration k, the clique is found by running a Grover's Search. After each iteration, the value of k is incremented until no clique is found. The author mentions the use of a binary search instead an iterative approach to speedup the algorithm by a $\log n$ factor. However, the author does not provide the implementation of the oracle for the Grover's Search. They do not even consider the complexity of the construction of such oracle; only its description is provided:

$$f(x) = \begin{cases} 1, & \text{if } x \text{ encodes a maximal clique with at least } k \text{ elements;} \\ 0, & \text{otherwise.} \end{cases}$$

Wie [6] proposed an algorithm based on Grover's Quantum Search for finding all maximal (not necessarily maximum) cliques in a graph G. The algorithm returns a superposition of all quantum states that describe a maximal clique in G. The author includes the description of the oracle, alongside with the instructions for creating the quantum circuit:

$$f(x) = \begin{cases} 1, & \text{if } x \text{ encodes a maximal clique;} \\ 0, & \text{otherwise.} \end{cases}$$

Since a maximum clique is a maximal clique, we can use Wie's oracle in Bojié's algorithm if the input state for Wie's oracle consists of a superposition of all states that encode a subgraph of G with at least k vertices.

In this paper we show how to prepare this input state for Wie's oracle, leading to an $\mathcal{O}(\sqrt{2^n}(n \log n)^2)$-time quantum algorithm for solving, with high probability, the Maximum Clique Problem. Remark that this is better than the $\mathcal{O}(3^{\frac{n}{3}})$ classical time, since

$$\lim_{n \to +\infty} \frac{\sqrt{2^n}(n \log n)^2}{3^{\frac{n}{3}}} = 0\,. \tag{1}$$

It is also observed that the performance gain, compared to the classical solution, is lower than the maximum square gain that Grover's algorithm can provide, since $\left(\sqrt{2^n}(n \log n)^2\right)^2 > 3^{\frac{n}{3}}$ for $n > 1$. Furthermore, our algorithm not only returns the cardinality of a maximum clique, but also a superposition of all quantum states which encode a maximum clique in the graph.

2. The algorithm

We present an algorithm to prepare a quantum state that is a superposition of all states that encode a clique candidate in G with at least k vertices, being k a fixed positive integer.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

A clique candidate C in a graph $G(V, E)$ is encoded as a binary string $|x\rangle = |x_1 x_2 ... x_n\rangle$ with $n = |V|$ qubits, wherein each qubit represents a vertex v_i of G and has the value:

$$|x_i\rangle = \begin{cases} |1\rangle, & \text{if } v_i \in C; \\ |0\rangle, & \text{otherwise.} \end{cases} \tag{2}$$

Note that a clique candidate with k vertices has exactly k qubits in the state $|1\rangle$. So, to find out the number of vertices in a clique candidate, we need to count the number of qubits in state $|1\rangle$. If this is done on a unary basis, the problem comes down to an ordering of the qubits in the binary string. In the unary basis we can also see if a clique candidate has at least k vertices by verifying if the k^{th} least significant qubit is in the state $|1\rangle$.

To count, on a unary basis, the number of qubits in the state $|x\rangle$ in a binary quantum string, we have developed a quantum algorithm based on the classical *insertion sort* algorithm. Substrings are iteratively ordered and expanded by inserting the remaining elements in the correct position. Each element inserted is either placed at the end of the string (most significant part), if it has value $|0\rangle$, or moved to the beginning (least significant part) if it has value $|1\rangle$. This quantum circuit can be constructed with CNOT and CSWAP gates, as described in Algorithm 1 and depicted in Figure 1.

Algorithm 1 Unary Qubit Counter

Input: $|x\rangle\,|a\rangle$; wherein $|a\rangle = |00\ldots0\rangle$
Output: $|x\rangle\,|u\rangle$; wherein $|u\rangle$ is a unary counting of the qubits of $|x\rangle$ in the state $|1\rangle$

1: **for** $i \in [1, 2, \ldots, n]$ **do**
2: apply a CNOT gate controlled by $|x_i\rangle$, targeting $|a_i\rangle$;
3: **end for**
4: **for** $i \in [2, 3, \ldots, n]$ **do**
5: **for** $j \in [i, i-1, \ldots, 2]$ **do**
6: apply a CSWAP gate controlled by $|x_i\rangle$, targeting $|a_j\rangle$ and $|a_{j-1}\rangle$;
7: **end for**
8: **end for**

The procedure for generating a superposition of all states with at least k vertices is based on another Grover's Quantum Search. This procedure aims to boost up the amplitude of the states in the superposition $|x\rangle$ that represent a clique candidate with at least k vertices. The oracle for this Grover's Search is depicted in Figure 2. The circuit is divided in 4 stages: *Copy to ancilla*, *Count in unary*, *Apply function* and *Clear ancilla*. The two first stages form the unary qubit counter (Figure 1). In the third stage we apply the function to the quantum state in a phase shift codification, as required by Grover's Algorithm, where the phase of the quantum input state $|x\rangle$ is inverted if $f(x) = 1$. This is done by applying a CNOT gate

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

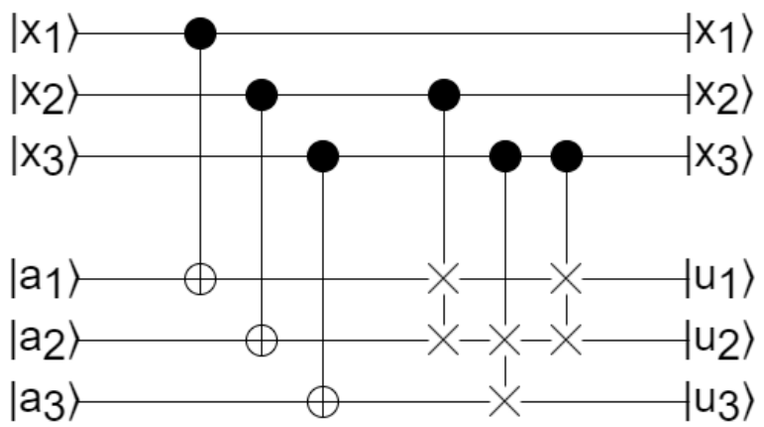


Figure 1: Unary qubit counter for $n = 3$

targeting a qubit in $|-\rangle$ state and controlled by a qubit $|c\rangle$ such that:

$$|c\rangle = \begin{cases} |1\rangle\,, & \text{if } f(x) = 1; \\ |0\rangle\,, & \text{otherwise.} \end{cases} \tag{3}$$

In this case, $|c\rangle$ is the k^{th} least significant qubit in the unary counting. In the last stage, the ancilla register is cleared by using a reversed circuit of the two first stages. Mathematically, the quantum circuit for this oracle O can be expressed as:

$$O\,|x\rangle = (-1)^{f(x)}\,|x\rangle\,. \tag{4}$$

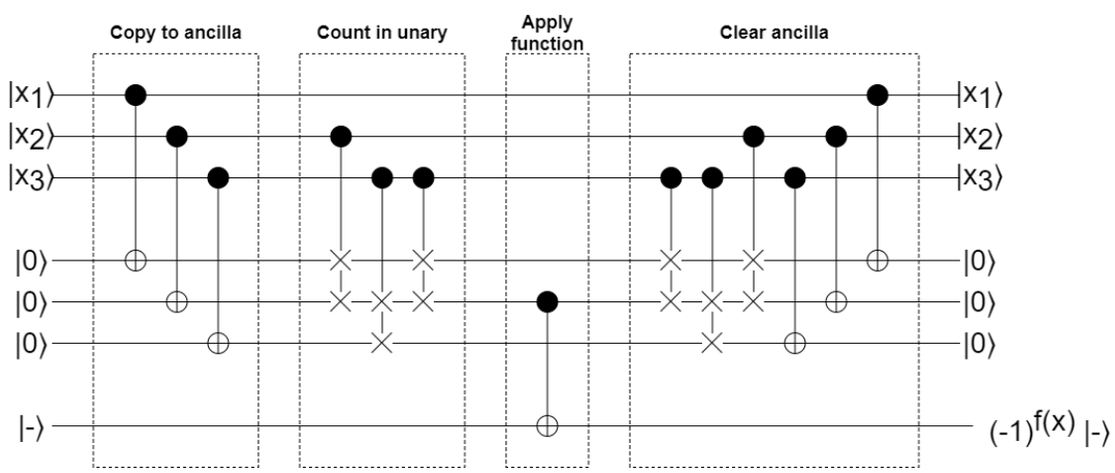


Figure 2: Quantum circuit for the oracle of the Grover's Search to generate the superposition of all clique candidates, with $n = 3$ and $k = 2$

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

The Grover's Search will leave the states that do not satisfy the oracle with a low amplitude. We can remove these states from the superposition by the use of quantum collapse and quantum entanglement phenomena. By adding another unary qubit counter at the end of the circuit and measuring the k^{th} least significant qubit of the counting, henceforth $|r\rangle$, the system will collapse, and the register $|x\rangle$ will be leaved in a superposition of all desired states if $|r\rangle = |1\rangle$, or an a superposition of all undesired states if $|r\rangle = |0\rangle$, according to the quantum entanglement phenomena. By Grover's effect, $|r\rangle$ will have a high probability of collapsing in $|1\rangle$, so we can repeat all the procedures until $|r\rangle$ collapses in $|1\rangle$ and $|x\rangle$ will be in the desired state of a superposition of all states with at least k qubits in $|1\rangle$. The full circuit for this algorithm is depicted in Figure 3.

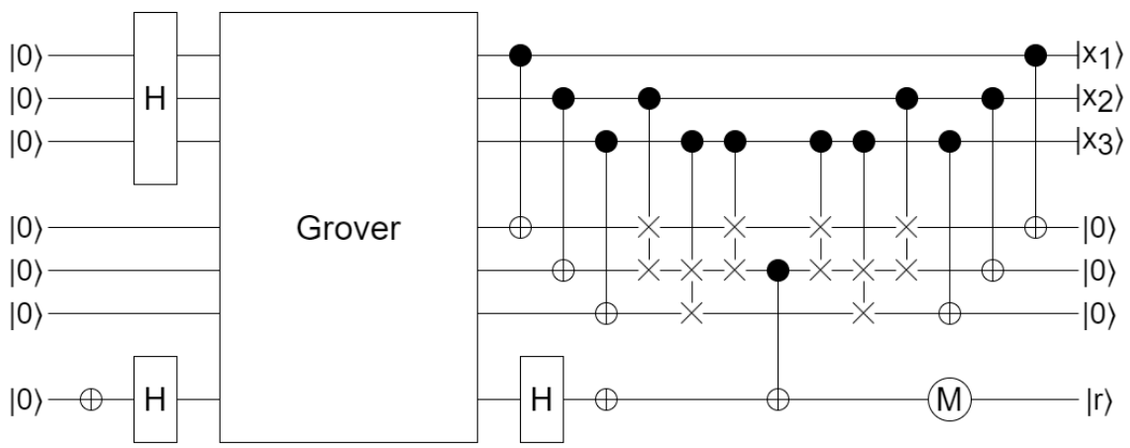


Figure 3: State generator for $n = 3$ and $k = 2$. The measurement is identified in the circuit by letter 'M'.

3. Brief analysis

Recall that our approach uses a binary search to make $\mathcal{O}(\log n)$ guesses for the value of k, building, for each guess, a circuit in $\mathcal{O}(n^2)$ time to prepare the input state for Wie's oracle through a Grover's Search, and then running another Grover's Search as in Wie's paper. This would yield an overall $\mathcal{O}(\sqrt{2^n}n^2 \log n)$ time complexity, but, to guarantee an overall $1 - \mathcal{O}(1/n)$ probability of error, each Grover's Search must be repeated some times.

Clearly, the probability p that our approach succeeds is bounded below by the probability that every Grover's Search individually and independently succeeds. Let p_i be the probability of success of the i^{th} Grover's Search, for $1 \leq i \leq 2\lceil \lg n \rceil$. To get $p \geq 1 - \mathcal{O}(1/n)$, as desired, it suffices to have each $p_i \geq (1 - \mathcal{O}(1/n))^{1/(2\lceil \lg n \rceil)}$, which we claim to be achieved by running each Grover's Search $\lfloor \lg n \rfloor$ times. The claim follows by observing that the probability that all these $\lfloor \lg n \rfloor$ iterations of Grover´s Search fail is $\mathcal{O}((1/n)^{\lg n})$, which is less than $1 - (1 - \mathcal{O}(1/n))^{1/(2\lceil \lg n \rceil)}$ for sufficiently large n. We have, then, our main result.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Theorem 1. *There is an $\mathcal{O}(\sqrt{2^n}(n \log n)^2)$-time and $\mathcal{O}(n^2)$-space quantum algorithm which, given an n-vertex graph G, outputs, with $1 - \mathcal{O}(1/n)$ probability of success, a state which is the superposition of all states which encode a maximum clique in G.* □

4. Final remarks

During the reviewing process, one of the reviewers pointed to us that a more efficient solution using Grover's algorithm for the Maximum Clique Problem could be achieved by an optimization algorithm such as the ones presented in [7] and [8]. For future work, we aim to investigate this reviewer's suggestion.

Acknowledgements

We thank the anonymous referee for the valuable suggestions given.

References

[1] R. M. Karp, Reducibility among combinatorial problems, Complexity of Computer Computations (1972) 85–103.

[2] J. Håstad, Clique is hard to approximate within $n^{1-\epsilon}$, Acta Math 182 (1999) 105–142. doi:10.1007/BF02392825.

[3] M. Nielsed, I. Chuang, Quantum Computation and Quantum Information, Cambridge University Press, 2000.

[4] R. Diestel, Graph Theory, Springer-Verlag, New York, 1997.

[5] A. Bojié, Quantum algorithm for finding a maximum clique in an undirected graph, JIOS 36 (2012) 91–98.

[6] C. R. Wie, A quantum circuit to construct all maximal cliques using Grover search algorithm, arXiv: Quantum Physics (2017). URL: `https://arxiv.org/abs/1711.06146`.

[7] C. Dürr, P. Høyer, A quantum algorithm for finding the minimum, 1996. URL: `https://arxiv.org/abs/quant-ph/9607014`.

[8] L. A. B. Kowada, C. Lavor, R. Portugal, C. M. H. Figueiredo, A new quantum algorithm for solving the minimum searching problem, International Journal of Quantum Information 6 (2008) 427–436.

On conformability of line graphs of complete graphs

Luerbio Faria[a], Mauro Nigro[a,b], Diana Sasaki[a]

[a] *Rio de Janeiro State University, Rio de Janeiro, RJ, Brazil*
[b] *Corresponding author: mauro.nigro@pos.ime.uerj.br*

Abstract

A Δ-regular graph G is conformable if it has a $(\Delta+1)$-vertex coloring where the cardinality of each vertex color class has the same parity as the order of the graph. A general characterization for conformable graphs is unknown. The importance of conformability is due to the fact that it can be an auxiliary tool toward determining the total chromatic number of a regular graph. Being conformable is a necessary condition for a graph to be *Type 1*. In this paper, we show a positive evidence to the conjecture proposed in 2018 which states that all line graphs of complete graphs $L(K_n)$ are *Type 1*, by proving that they are all conformable.

Keywords: conformable, total coloring, complete graph, line graph

1. Introduction

Let $G=(V,E)$ be a simple connected graph. A *k-total coloring of G* is an assignment of k colors to the vertices and edges of G so that adjacent or incident elements have different colors. The *total chromatic number of G*, denoted by $\chi''(G)$, is the smallest k for which G has a k-total coloring. Clearly, $\chi''(G) \geq \Delta(G)+1$ and the Total Coloring Conjecture (TCC) states that the total chromatic number of any graph is at most $\Delta(G)+2$, where $\Delta(G)$ is the maximum degree of the graph [1, 2]. If the TCC holds, graphs with $\chi''(G)=\Delta(G)+1$ are called *Type 1*, and graphs with $\chi''(G)=\Delta(G)+2$ are called *Type 2*. In 1989, Sánchez-Arroyo [3] proved that determining the total chromatic number of an arbitrary graph is a NP-hard problem.

A *Type 1* graph has a nice structural property due to Chetwynd and Hilton [4]. Let the *deficiency of G* be $def(G)=\sum_{v\in V(G)}(\Delta(G)-d_G(v))$, where $d_G(v)$ is the degree of a vertex v in G. A vertex coloring $\varphi: V(G)\rightarrow\{1,2,...,\Delta(G)+1\}$ is called *conformable* if the number of color classes (including empty color classes) of parity different from that of $|V(G)|$ is at most $def(G)$. Note that since regular graphs have deficiency 0, each color class has the same parity as $|V(G)|$. A graph is conformable if and only if it has a conformable vertex coloring, since otherwise it is called non-conformable.

It is well known that every *Type 1* graph is conformable [4]. In the same paper, Chetwynd and Hilton conjectured that for graphs with maximum degree greater than one half of their order the converse is essentially true. However, in some cases this is not true: For instance,

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

the complete bipartite graphs $K_{n,n}$, for even $n > 1$, and the Möbius ladders M_{2n}, for $n > 3$, are conformable and *Type 2* [4]. Determining suitable conformable vertex colorings or proving the non-conformability of a graph G are usefull tools to determine the total chromatic number of a graph. Indeed, every non-conformable graph is *Type 2* and suitable conformable vertex colorings may be extended to $(\Delta(G) + 1)$-total colorings.

There are few results on conformability, such as a characterization of non-conformability of some graphs and the verification of Chetwynd and Hilton's conjecture for special classes of graphs [5, 6]. In this paper, we prove that the line graphs of complete graphs $L(K_n)$ are conformable, contributing to the conjecture proposed by Vignesh et al. [7] in 2018 that all line graphs $L(K_n)$ are *Type 1*. Note that the small cases $L(K_3) \simeq K_3$ and $L(K_4) \simeq C_6^2$ are trivial examples of *Type 1* graphs.

2. Main Result

In this section, we prove that all graphs $L(K_n)$ are conformable. Note that $V(L(K_n)) = \frac{n(n-1)}{2}$, and first we mention a usefull lemma.

Lemma 1. *If G is regular of degree k, then $L(G)$ is regular of degree $2k - 2$.*

Consequently, the degree of $L(K_n)$ is $2(n-1) - 2 = 2n - 4$ and so in order to prove that $L(K_n)$ is conformable, we exihibt a vertex coloring with $\Delta(L(K_n)) + 1 = 2n - 3$ colors such that the number of vertices in each color class has the same parity as $|V(L(K_n))|$.

Theorem 1. *The graphs $L(K_n)$ are conformable.*

Proof. First consider $V(K_n) = \{v_0, v_1, ..., v_{n-1}\}$. By definition of line graphs, each edge of K_n is identified with a vertex in $L(K_n)$.

The proof is separated into two cases as follows:

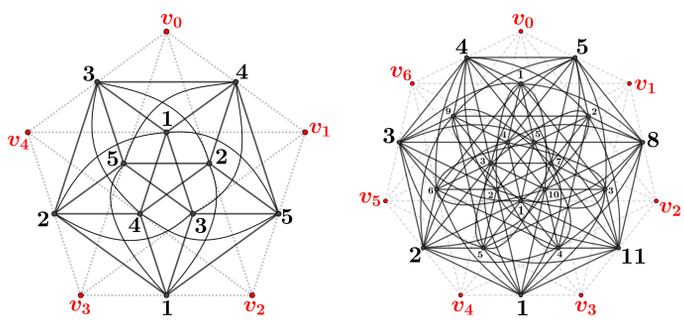
- For n odd: let $n = 2k - 1$, the number of vertices of $L(K_n)$ is given by $\frac{n(n-1)}{2} = (2k-1)(k-1)$. So, we consider the cases where k is odd and k is even, separately.

  If k is odd, then the number of vertices of $L(K_n)$ is even. Therefore, for each $1 \leq p \leq n$, we will define the set C_p as the set of vertices of $L(K_n)$ with color p. This is given by $C_p = \left\{v_{p-1-q}v_{p-1+q} \mid 1 \leq q \leq \frac{n-1}{2}\right\}$, with $v_i \in V(K_n)$ where the indexes of the vertices are considered modulo n. Note that C_p is a maximum matching in K_n, and so it is a maximal independent set on $L(K_n)$. Moreover, $\bigcup_{i=1}^{n} C_i = E(K_n) = V(L(K_n))$. Consequently this vertex coloring is conformable, because $|C_p| = \frac{n-1}{2} = k - 1$ is even and we use $\Delta(L(K_n)) + 1$ colors (where $n-3$ color classes are empty). Figure 1a shows an example of the vertex coloring for this case.

  If k is even, then the number of vertices of $L(K_n)$ is odd. We use C_p as before, for $1 \leq p \leq n-2$. The vertices $v_{n-2-q}v_{n-2+q}$ and $v_{n-1-q}v_{n-1+q}$ will be assigned with colors $n - 1 + q - 1$ and $n + q - 1 + \frac{n-1}{2} - 1$, respectively for each $1 \leq q \leq \frac{n-1}{2}$. As

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

$|C_p| = \frac{n-1}{2} = k-1$ is odd, the colors assigned to $v_{n-2+q}v_{n-2-q}$ and $v_{n-1+q}v_{n-1-q}$ appear only once and we use $n-2+\frac{n-1}{2}+\frac{n-1}{2} = 2n-3 = \Delta(L(K_n))+1$ colors. So, we conclude that this vertex coloring is conformable. Figure 1b shows an example of the vertex coloring for this case.



(a) Conformable vertex coloring for $L(K_5)$

(b) Conformable vertex coloring for $L(K_7)$

Figure 1: A depiction of a conformable vertex coloring for $L(K_n)$, with n odd.

- For n even: let $n = 2k$, note that the number of vertices of $L(K_n)$ is given by $\frac{n(n-1)}{2} = k(2k-1)$. Therefore, we consider the cases where k is odd and k is even, separately.

  If k is odd, then the number of vertices of $L(K_n)$ is odd. So, we define the set A_p as the set of vertices of $L(K_n)$ with color p. It is given by $A_p = \left\{v_{p-q}v_{p-1+q} \mid 1 \leq q \leq \frac{n}{2}\right\}$, with $1 \leq p \leq \frac{n}{2}$. The Figure 2 shows the set A_p for $L(K_{10})$ by edge coloring of K_{10}. Note that A_p is a maximum matching of K_n for each p. Moreover, $K_n \setminus \bigcup_{p=1}^{\frac{n}{2}} A_p$ are two copies of $K_{\frac{n}{2}} \simeq K_k$. That is, the subgraph induced by the vertices of odd indexes is the graph K_k, as well as the subgraph induced by the vertices of even indexes is the other graph K_k. Let $u_i = v_{2i}$ and $w_i = v_{2i+1}$, with $0 \leq i \leq \frac{n}{2}-1$. Consider the set of color $j = \frac{n}{2}+p$ as $B_j = \{u_{j-1-q}u_{j-1+q} \mid 1 \leq q \leq \frac{k-1}{2}\} \cup \{w_{j-1-q}w_{j-1+q} \mid 1 \leq q \leq \frac{k-1}{2}-1\}$, with $1 \leq p \leq k-1$ and the indexes of vertices are considered modulo k. Figure 2 shows an example of the construction of a vertex coloring using B_j for $L(K_{10})$. The remaining $2k-2$ uncolored edges of each K_k will receive different colors from the ones used until now, i.e., each one of these edges is an element of a different color class.

  We used $k+k-1+2k-2 = 4k-3 = 2n-3 = \Delta(L(K_n))+1$ colors. Moreover, the parity of each class of color $|A_p| = k$ is odd as well as $|B_j| = \frac{k-1}{2}+\frac{k-1}{2}-1 = k-2$, and finally the remaining color classes, which have only one element, have parity odd. Therefore, we conclude that this vertex coloring is conformable.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

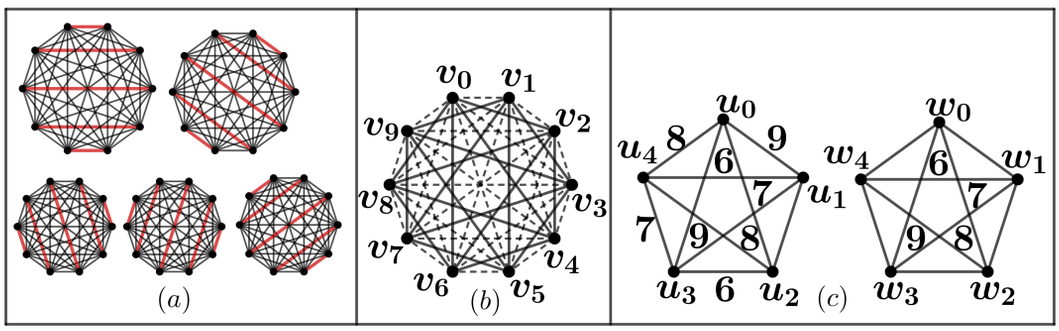


Figure 2: (a) The set A_p to K_{10} for each $1 \leq p \leq 5$, represented by edge with red color; (b) the induced subgraph that are two copies of K_5; (c) and the set B_j to 2 copies of K_5, for each $6 \leq j \leq 9$.

If k is even, then the number of vertices of $L(K_n)$ is even. First we consider A_{p_0}, defined as above with $1 \leq p_0 \leq \frac{n}{2}$. The Figure 3 shows the set A_{p_0} for $L(K_8)$ by edge coloring of K_8. Similarly, $K_n \setminus \bigcup_{p_0=1}^{\frac{n}{2}} A_{p_0}$ results in two copies of K_k. As k is even, so $k = 2k_1$. If k_1 is even we can repeat, recursively, this process again, using $A_{\frac{n}{2}+p_1}$ in each K_{k_1}, where $1 \leq p_1 \leq \frac{n}{4}$ (Figure 3) and end when for some k_j we have $K_{k_j} \setminus \bigcup_{p_j=1}^{\frac{n}{2^j}} A_{\sum_{i=1}^{j-1} \frac{n}{2^i}+p_j} = \emptyset$. If at any time we have k_j odd, then we apply the coloring C_ℓ, with $\ell = \left(\frac{n}{2} + \frac{n}{4} + \cdots + \frac{n}{2^j}\right) + p_j$ and $1 \leq p_j \leq k_j - 1$, for each copy K_{k_j}. Note that, for any color λ used in A_λ, $|A_\lambda|$ is even and in every process $|C_\ell|$ is used for even copies. Therefore, the parity of each color class is the same to the number of vertices and we use at most $\frac{\left(\frac{n}{2}\right)\left(1-\left(\frac{1}{2}\right)^{\lfloor \log_2 n \rfloor}\right)}{1-\frac{1}{2}} = n\left(1-\left(\frac{1}{2}\right)^{\lfloor \log_2 n \rfloor}\right)$ colors, with $n\left(1+\left(\frac{1}{2}\right)^{\lfloor \log_2 n \rfloor}\right) - 3$ color classes with zero occurrences. We conclude that this vertex coloring is conformable.

□

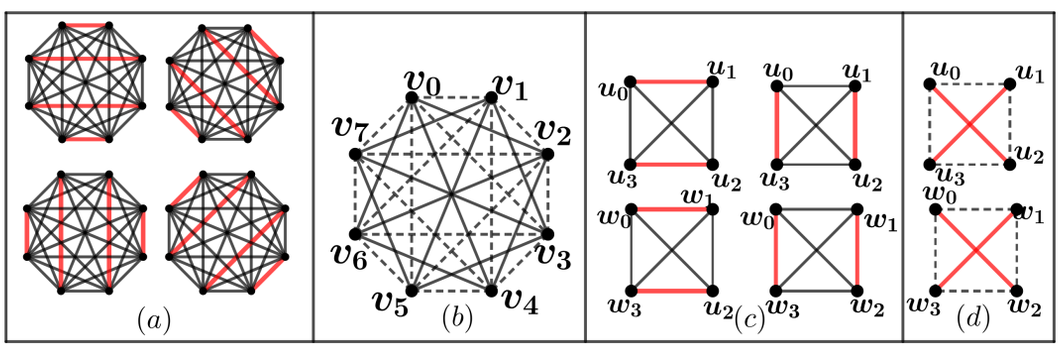


Figure 3: (a) The set A_{p_0} to K_8, for each $1 \leq p_0 \leq 4$, represented by edge with red color; (b) the induced subgraph that are two copies of K_4; (c) the set $A_{\frac{n}{2}+p_1}$ to 2 copies of K_4, for each $5 \leq p_1 \leq 6$; (d) and the set $A_{\frac{n}{2}+\frac{n}{4}+p_2}$ to 4 copies of K_2, for $p_2 = 7$.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

3. Conclusion

In this paper, we proved that all line graphs $L(K_n)$ are conformable. Note that we can easily extend our conformable vertex coloring of $L(K_5)$ to a total coloring of this graph, showing that $L(K_5)$ is *Type 1* (Figure 4). This is a single and preliminar evidence that the conjecture proposed by Vignesh et al. [7] is valid and we will continue to investigate this conjecture.

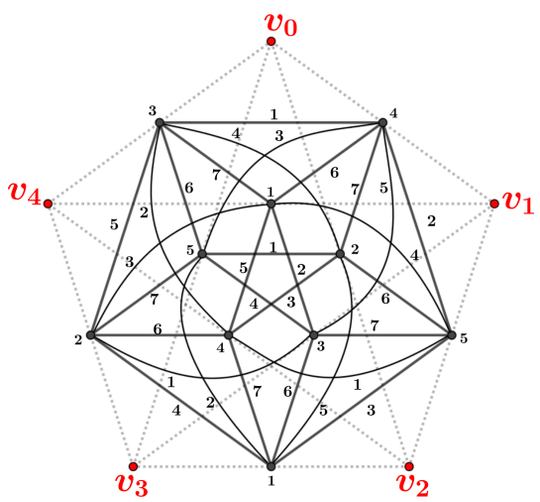


Figure 4: A $(\Delta(L(K_5)) + 1)$-total coloring to $L(K_5)$.

References

[1] M. Behzad, Graphs and and their chromatic numbers, Ph.D. thesis, Michigan State University, 1965.

[2] V. Vizing, On an estimate of the chromatic class of a p-graph, Metody Diskret. Analiz. (1964) 25–30.

[3] A. Sánchez-Arroyo, Determining the total colouring number is NP-hard, Discrete Math. (1989) 315–319.

[4] A. G. Chetwynd, A. J. W. Hilton, Some refinements of the total chromatic number conjecture, Congr. Numer. (1988) 195–216.

[5] G. M. Hamilton, A. J. W. Hilton, H. R. F. Hind, Totally critical even order graphs, J. Comb. Theory Ser. B. 76 (1999) 262–279.

[6] A. J. W. Hilton, H. R. Hind, Non-conformable subgraphs of non-conformable graphs, Discrete Math. (2002) 203–224.

[7] R. Vignesh, J. Geetha, K. Somasundaram, Total coloring conjecture for certain classes of graphs, Algorithms 11 (2018).

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Mapeamento Sistemático sobre Redes Neurais Convolucionais Aplicadas na Segmentação de Imagens Aéreas

Fabricio Bizotto[1], Mauren Louise Sguario Coelho De Andrade[1]

[a]*Departamento Acadêmico de Informática*
Programa de Pós-Graduação em Ciência da Computação (PPGCC)
Universidade Tecnológica Federal do Paraná (UTFPR)
Ponta Grossa, Brasil
[b]*fabriciobizotto@alunos.utfpr.edu.br, mlsguario@utfpr.edu.br*

Resumo

Existem muitos estudos atuais publicados sobre Redes Neurais Convolucionais (CNN) em diferentes contextos. Foi feito um mapeamento sistemático sobre a utilização de redes CNN aplicadas na segmentação de imagens aéreas. A partir do levantamento inicialmente feito, 5 artigos foram selecionados e analisados. Como resultado, percebe-se que essa linha de pesquisa vem crescendo, mas seria necessário analisar mais publicações para responder melhor às questões de pesquisa. Entre as redes CNN utilizadas, cada uma conseguiu contribuir de alguma forma na segmentação de imagens aéreas.

Abstract

There are many current studies published on Convolutional Neural Networks (CNN) in different contexts. A systematic mapping of the use of CNN networks applied in the segmentation of aerial images was carried out. From the survey initially carried out, 5 articles were selected and analyzed. As a result, it is clear that this line of research has been growing, but it would be necessary to analyze more publications to better answer the research questions. Among the CNN networks used, each managed to contribute in some way to the segmentation of aerial images.

Palavras-chaves: Rede Neural Convolucional, Imagens Aéreas, Segmentação Semântica

1. Introdução

O desenvolvimento de novas abordagens em Machine Learning (ML) proporciona aos pesquisadores alternativas para a resolução de problemas gerais, através de modelos computacionais criados e treinados a partir de uma grande quantidade de dados. Uma das abordagens que está ganhando espaço é a Rede Neural Convolucional (CNN), oriunda das áreas de processamento de imagens e visão computacional. O algoritmo possibilita vantagens em comparação aos métodos anteriores de última geração, superando nas tarefas de

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

classificação de imagens, reconhecimento e localização de objetos e rotulagem semântica de pixels[1].

A primeira implementação da CNN foi feita por LeCun [2]. O experimento foi aplicado no reconhecimento de dígitos manuscritos e resultou em uma precisão de 99,2% de acerto na identificação dos números. O processamento de imagens utilizando CNN cresceu nos anos seguintes. São utilizadas em diversos contextos. No contexto relacionado a imagens aéreas, as redes CNN são utilizadas, por exemplo, para o mapeamento aéreo de piscinas [3], na avaliação dos danos sísmicos em edifícios, para melhorar a resposta a emergências em locais sujeitos a terremotos [4], para detectar objetos de interesse em imagens aéreas capturadas por UAVs [5] e também para identificar e classificar ervas daninhas que crescem no meio do soja [6]. Grandes repositórios de imagens tais como ImageNet surgiram e a otimização no desempenho das unidades de processamento gráfico (GPU) contribuíram para o sucesso das CNN. Além disso, competições como ImageNet Challenge [7] e Visual Object Classes Challenge [8], publicadas na revista International Journal of Computer Vision, atraem pesquisadores. Resultados desse tipo de competição gerou modelos de rede CNN, tais como AlexNet [9] e VGG-Net [10].

Com isso, a segmentação de imagens, cujo objetivo é separar a área/objeto de interesse de seu fundo, sofreu grande avanço com a aplicação de segmentação semântica. A segmentação semântica é caracterizada por classificar objetos que possuem algum rótulo, conhecido ou desconhecido. [11] Essa técnica pode ser utilizada em diferentes contextos, desde o sensoriamento remoto por satélite [12], análise de vídeos de embriões em desenvolvimento [13] e interação humano-computador [11].

Dentre as áreas de aplicação de segmentação semântica citadas, a segmentação e rotulagem de imagens aéreas tem possibilitado o acompanhamento e prevenção de problemas relacionados ao desmatamento e apropriação de áreas de preservação em florestas em diferentes regiões do mundo [14]. Neste sentido, o desenvolvimento de uma nova metodologia de segmentação semântica com uso de CNN capaz de aprimorar o reconhecimento e rotulagens de áreas de preservação poderá contribuir com o trabalho das autoridades ambientais na fiscalização e demarcação de áreas já afetadas.

Portanto, com o objetivo de responder algumas questões relacionadas à aplicação de CNN em imagens aéreas, este trabalho apresenta um mapeamento sistemático, o qual irá identificar e mapear quais abordagens têm sido utilizadas na segmentação semântica de imagens aéreas com o uso de CNN. A ideia principal é coletar informações a fim de identificar estratégias para o aprimoramento dos resultados de classificação e segmentação de áreas de preservação brasileiras, em especial na área de preservação da Mata Atlântica na região de Petrópolis, Rio de Janeiro. Este artigo está estruturado da seguinte maneira: Após esta breve introdução, na seção 2 é descrito o método de pesquisa. Na seção 3, os resultados são discutivos e na seção 4 uma breve conclusão.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

2. Método de Pesquisa

Para levantar o estado da arte sobre a utilização de redes CNN no contexto de imagens aéreas, foi feito um mapeamento sistemático seguindo os procedimentos essenciais propostos por Petersen [15], as quais serão descritos a seguir:

2.1. Metodologia

O mapeamento foi conduzido entre abril e maio de 2021 e cobriu publicações sobre utilização de redes CNN em imagens aéreas, no período de 2017 a 2021.

2.2. Questões de Pesquisa

Foram definidas questões que considerem pontos relevantes a serem estudados:

- QP1: Quais e como as redes CNN são utilizadas para classificação de imagens aéreas?
- QP2: Em quais contextos as redes CNN foram aplicadas?
- QP3: Quais redes CNN funcionam melhor para classificação de imagens aéreas?

2.3. Palavras chave e string de busca

As palavras chave estão em inglês e foram definidas de acordo com o tema pesquisado. São elas: CNN, Convolutional Neural Network, Image Segmentation, Image Classification, Aerial Image e Aerial Imagery. A string de busca foi criada a partir da combinação lógica entre elas, e ficou da seguinte forma: (*"CNN" OR "Convolutional neural network"*) AND (*!Image Segmentation" OR "Image Classification"*) AND (*"Aerial image" OR "Aerial Imagery"*).

2.4. Critérios de inclusão e exclusão

Para refinar a pesquisa, os seguintes critérios de inclusão de artigos foram aplicados: (i) artigos escritos em inglês; (ii) artigos com relevância ao assunto abordado; (iii) revistas com relevância na área; (iv) artigos com qualis maior ou igual a B1. Também foram aplicados os seguintes critérios de exclusão: (i) artigos duplicados; (ii) artigos com título e resumo fora do assunto abordado.

2.5. Bases de Pesquisa

As máquinas de busca utilizadas foram: *ACM Digital Library, IEEE Xplore, Scopus e Science Direct*. A string de busca utilizada foi a mesma em todas as bases. A pesquisa resultou em 837 artigos não duplicados. A maior parte dos estudos é recente e sugere um crescente interesse pelo tema nos últimos anos, devido a fatores como avanço em termos de hardware e economia de tempo, aliados a resultados promissores em termos de segmentação. Foram selecionados 5 artigos para serem analisados.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

3. Resultados e Discussão

Respondendo a QP1, cada aplicação faz o pré-processamento da rede, normalizando os dados de entrada e separando as imagens em grupos para a fase de treinamento, validação e teste. Cada estrutura de rede CNN é configurada com uma quantidade diferente de camadas de convolução. A Figura 1 mostra os *datasets* e as redes CNN utilizadas em cada trabalho.

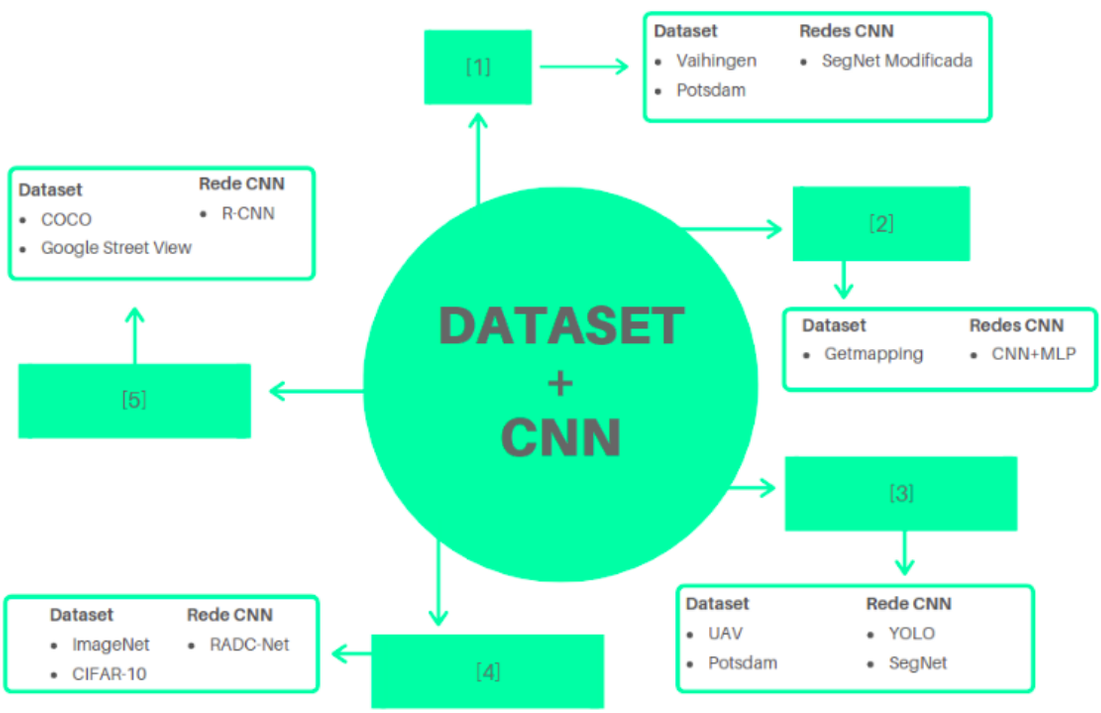


Figura 1: *Datasets* e redes CNN utilizadas

Cada estudo utilizou uma rede CNN diferente e *datasets* com imagens aéreas. O *dataset* mais utilizado foi da cidade de Potsdam, Alemanha. Respondendo a QP2, diferentes contextos foram abordados nos trabalhos: prevenção de risco de incêndio; área agrícola; mapeamento de árvores. As questões anteriores colaboram para responder a QP3. Contudo, não é possível definir qual rede CNN funciona melhor para classificar imagens aéreas. Cada implementação auxiliou de alguma forma, seja para reduzir o erro de validação, para reduzir problemas de overfitting ou para identificação de árvores com alta taxa de precisão.

4. Conclusão

O foco deste estudo de mapeamento sistemático é investigar como as redes CNN estão sendo aplicadas para segmentação semântica no contexto de imagens aéreas. Apesar de terem sido mapeados 837 artigos, apenas 5 foram utilizados. Mais artigos precisam ser

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

considerados para enriquecer o mapeamento e obter mais dados para responder melhor às questões de pesquisa elaboradas. O que se conclui desse mapeamento sistemático é que a grande quantidade de publicações recentes em jornais e revistas demonstra o crescente interesse pelo tema.

Referências

[1] K. Nogueira, O. A. B. Penatti, A. Santos, Towards better exploiting convolutional neural networks for remote sensing scene classification, Pattern Recogn. 61 (2017) 539–556.

[2] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.

[3] B. Lima, L. Ferreira, J. Moura, Helping to detect legal swimming pools with deep learning and data visualization, Procedia Computer Science 181 (2021) 1058–1065.

[4] C. Xiong, Q. Li, X. Lu., Automated regional seismic damage assessment of buildings using an unmanned aerial vehicle and a convolutional neural network, Automation in Construction 109 (2020) 102994.

[5] L. Zhaokun, L. X. Zhao, Y. Liu, B. Huang, Z. H. Richang, A lightweight multi-scale aggregated model for detecting aerial images captured by uavs, Journal of Visual Communication and Image Representation 77 (2021) 103058.

[6] A. S. Ferreira, D. M. Freitas, G. G. Silva, H. Pistori, M. T. Folhesc., Weed detection in soybean crops using convnets, Computers and Electronics in Agriculture 143 (2017) 314–324.

[7] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, C. Berg, L. Fei-Fei, Imagenet large scale visual recognition challenge, Int. J. Comput. Vision 115 (3) (2015) 211–252.

[8] M. Everingham, S. M. A. Eslami, L. V. Gool, C. K. I. Williams, J. Winn, A. Zisserman, The pascal visual object classes challenge: A retrospective, Int. J. Comput. Vision 111 (1) (2015) 98–136.

[9] A. Krizhevsky, I. Sutskever, G. E. Hinton., Imagenet classification with deep convolutional neural networks, Advances in neural information processing systems 25 (2012) 1097–1105.

[10] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, Computer Vision and Pattern Recognition (2015).

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

[11] G. Yanming, L. Yu, T. Georgiou, M. S. Lew, A review of semantic segmentation using deep neural networks, International Journal of Multimedia Information Retrieval 7 (2018) 87–93.

[12] M. Coffe, B. Schaeffer, R. Zimmerman, V. Hill, J. Li, K. Islam, P. Whitman, Performance across worldview-2 and rapideye for reproducible seagrass mapping, Remote Sensing of Environment 250 (2020) 112036.

[13] F. Ning, D. Delhomme, Y. LeCun, F. Piano, L. Bottou, P. Barbano, Toward automatic phenotyping of developing embryos from videos, IEEE Transactions on Image Processing 14 (9) (2005) 1360–1371.

[14] A. A. Sant'Anna, L. Costa, Environmental regulation and bail outs under weak state capacity: Deforestation in the brazilian amazon, Ecological Economics 186 (2021) 107071.

[15] K. Petersen, S. Vakkalanka, L. Kuzniarz, Guidelines for conducting systematic mapping studies in software engineering, Information and Software Technology 64 (2015) 1–18.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Desenvolvimento de um Algoritmo de Classificação de Dados Baseado na Relevância da Informação

Júlio César Guimarães Costa[a,b], Gleifer Vaz Alves[a], André Pinz Borges[a]

[a] *Grupo de Pesquisa, LaCA-IS, Departamento Academico de Informática, Universidade Técnologica Federal do Paraná (UTFPR), R. Doutor Washington Subtil Chueire, 330, Paraná, Brasil*
[b] *Autor para correspondência: juliocesargcosta123@gmail.com*

Resumo

Algoritmos de classificação são uma ferramenta poderosa, principalmente nos dias de hoje, quando os dados estão se tornando cada vez mais valiosos. Este artigo propõe um algoritmo chamado Prevh que baseia-se em uma nova forma de se visualizar as informações. Para tal foi escolhida uma metodologia qualitativa, básica e explicativa. Depois de ser desenvolvido o algoritmo Prevh, foi submetido a uma série de comparações com o algoritmo de classificação KNN o qual mais se assemelha. Uma delas pode ser inclusive observada na seção 4 deste artigo. Por fim foi observado que além de considerar a relevância da informação, o algoritmo Prevh foi capaz de ser conclusivo em situações nas quais o algoritmo KNN não foi.

Abstract

Classification algorithms are a powerful tool, especially today, when data is becoming increasingly valuable. This paper proposes an algorithm called Prevh, which is based in a new way of visualize information. For this purpose, a qualitative, basic and explanatory methodology was chosen. Initially, we have developed the algorithm, next we have run a series of comparisons between Prevh and KNN classification algorithm, since both share some similarities. In Section 4, we describe such similar features. Finally, we have observed that in addition to considering the information relevance, Prevh was able to be conclusive in some scenarios where KNN was not.

Palavras-chaves: Aprendizagem de Maquina, Mineração de Dados, Classificação de Dados

1. Introdução

Aplicações e tecnologias que envolvem problemas de classificação são mais comuns do que podemos imaginar nos dias de hoje, e estão presentes em uma infinidade de algoritmos que envolvam Aprendizagem de Máquina em geral [1]. Alguns exemplos podem ser notados em algoritmos de plataformas como Facebook, Amazon, Google, entre outros. Estas empresas trabalham com uma grande quantidade de dados e se utilizam dessas importantes ferramentas em traçamentos de perfil de clientes, classificação de categorias, entre outras aplicações [2].

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Neste artigo é proposto o algoritmo Prevh que se mostrou uma ferramenta poderosa quando utilizada em conjunto com um outro algoritmo de classificação já consolidado, chamado KNN. Ambos algoritmos têm o objetivo de classificar uma entidade X que hipoteticamente surja em um espaço n-dimensional definido por K entidades. Será possivel observar que as situações de aplicação do algoritmo Prevh ocorrem principalmente em situações em que o algoritmo KNN é incerto quanto a classificação, o que será demonstrado e comprovado no decorrer do artigo.

Além da aplicação do algoritmo a relevância da informação pode ser definida como um valor ou peso que é calculado e atribuído tanto ao dado de treinamento quanto no dado que se pretende classificar, e que representa justamente a veracidade da informação, este recurso pode ser e será utilizado na previsão de diferentes cenários sequenciais definidos por processos estocásticos [3] que serão explorados em trabalhos futuros.

2. Definição do procedimento matemático

2.1. Terminologia

Inicialmente, apresenta-se a terminologia de alguns elementos usados neste trabalho.

Cluster C = Agrupamento de pontos representantes de um mesmo resultado.
Entidade = Ponto de coordenadas $(x,y,\ldots,n)$ posicionado no espaço n-dimensional.
Relevância = O quanto uma entidade representa um determinado *cluster*, variando de 0 a $\sqrt{nD}$, onde nD é o número de dimensões do sistema.
Entidade X = Entidade que se pretende classificar.
Entidade a = Entidade representante de um *cluster* C, proveniente dos dados de entrada.

2.2. Introdução ao procedimento matemático

Buscando facilitar o entendimento do conceito matemático abordado neste artigo é interessante que seja feita uma breve referência ao seguinte conceito - "Dois corpos atraem-se por uma força que é diretamente proporcional ao produto de suas massas e inversamente proporcional ao quadrado da distância que os separa" [4], partindo de tal conceito foi feita uma associação entre a massa e a informação, e consequentemente a força de atração e a relevância, foi então definido que a relevância da informação é diretamente proporcional ao índice de pertinência de uma entidade a um determinado *cluster*, e portanto ambas são inversamente proporcionais a distância que as separa de outras entidades.

Para a aplicação do algoritmo, é estritamente necessário que o universo estudado (incluso a entidade X) seja normalizado entre os valores 0 e 1, para que seja possível calcular a maior distância euclidiana no espaço. Tal regra pode ser observada com a aplicação do teorema de Pitágoras no espaço n-dimensional [5], e este valor será utilizado como referência para a maior relevância que uma instância pode ter no universo estudado.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

3. Fórmulas matemáticas

A partir dos conceitos apresentados foram definidas as fórmulas a seguir:

Definição 1 (Média Ponderada da soma das Distâncias Euclidianas)**.**

$$Mdp^{K}_{X\to eC} = \frac{\sum_{a=1}^{neC}\sqrt{\sum_{i=1}^{nD}(Cx - Ca)^2} * R_a}{neC} \tag{1}$$

onde:

$Mdp^{K}_{X\to eC}$ *= Média Ponderada da soma das Distâncias Euclidianas.*
neC *= Número de entidades representantes do cluster* C*.*
a *= entidade do cluster* C*.*
nD *= Número de dimensões do sistema.*
Cx *= Coordenada da entidade* X*.*
Ca *= Coordenada da entidade* a*.*
Ra *= Relevância da entidade* a *que é definida nos dados de treinamento baseando-se na veracidade da informação.*

A média ponderada da soma das distâncias euclidianas entre a entidade X *e as entidades* A *representantes de um cluster* C *em um espaço delimitado por* K*, é calculada pelo somatório da distância euclidiana entre* X *e* A *representante de* C*, multiplicado por* Ra*. Tudo dividido pelo número de entidades de* C*.*

Definição 2 (Índice de pertinência de X em C no espaço K)**.**

$$P^{K}_{X\to C} = 1 - \frac{Mdp^{K}_{X\to eC}}{\sum_{i=1}^{neK} Mdp^{K}_{X\to eC}} \tag{2}$$

onde,

$P^{K}_{X\to C}$ *= Índice de Pertinência de* X *em* C *no espaço* K*.*
$Mdp^{K}_{X\to eC}$ *= Média Ponderada das Distâncias Euclidianas.*
neK *= Número de entidades no espaço delimitado por* K*.*

O índice de pertinência de uma entidade X *em* C *em um espaço delimitado por* K*, é definida por 1 menos a Média Ponderada da soma das Distâncias Euclidianas sobre a soma de todas as MDPs entre a entidade* X *e todas entidades no espaço delimitado por* K*.*

Definição 3 (Calculo da relevância de X em C no espaço K)**.**

$$R^{K}_{X\to C} = \sqrt{nD} * P^{K}_{X\to C} \tag{3}$$

onde,

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

nD = *Número de dimensões do sistema.*
$P^K_{x \to C}$ = *Índice de Pertinência de* X *em* C *no espaço* K.
$R^K_{x \to C}$ = *Relevância de* X *em* C *no espaço* K.

A relevância da entidade X *para um determinado cluster* C *em um espaço delimitado por* K*, é definida pela multiplicação entre a raiz do número de dimensões* nD *pelo índice de pertinência de* X *em* C *no espaço* K.

4. Comparativo entre o algoritmo de classificação KNN e o algoritmo Prevh

Foi proposto o seguinte teste comparativo: em um sistema de duas dimensões normalizado entre 0 e 1, pretende-se classificar uma entidade X de coordenadas (0.5,0.5), para um espaço delimitado por $K = 3$, e para tal foram aplicados, primeiramente o algoritmo KNN e posteriormente o algoritmo Prevh, após os resultados obtidos foram comparados. Os dados de entrada são apresentados na Tabela 1 e Figura 1.

Ca_x	Ca_y	Ca_C	Relevância (Ra)	Distância Euclidiana
0,1	0,1		1,41	0,56
0,1	0,2		1,41	0,5
0,4	0,4		1,41	0,14
0,4	0,9		1,41	0,41
0,5	0,8		1,41	0,3
0,5	0,6		1,41	0,1
0,6	0,4		1,41	0,14
0,8	0,3		1,41	0,36
0,9	0,7		1,41	0,44

Tabela 1: Representação descritiva dos dados de teste incluindo a distância euclidiana já calculada entre as instâncias e a instância que se pretende classificar X(0.5,0.5), com os K=3 vizinhos mais próximos marcados em amarelo.

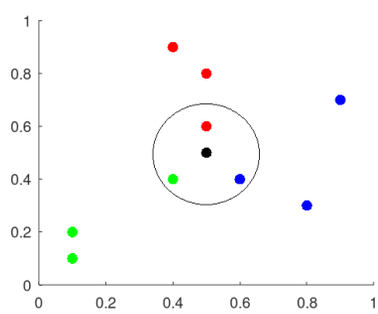


Figura 1: Representação gráfica dos dados de teste com a área de atuação limitada a K=3.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

4.1. Aplicando o algoritmo KNN

Ao ser aplicado, o algoritmo KNN [6] para K = 3 foram obtidos os seguintes resultados:
0,33 ou 33% de chance de X ser uma entidade da cor azul.
0,33 ou 33% de chance de X ser uma entidade da cor vermelha.
0,33 ou 33% de chance de X ser uma entidade da cor verde.

4.2. Aplicando o algoritmo Prevh

Ao serem aplicadas as fórmulas matemáticas descritas anteriormente neste artigo, os resultados ilustrados na Tabela 2 foram obtidos.

C_c	$Mdp^K_{X\to eC}$	$P^K_{X\to C}$	$R^K_{X\to C}$
	0,1974	0,6316	0,8905
	0,141	0,7369	1,0390
	0,1974	0,6316	0,8905

Tabela 2: Resultados obtidos pelo algoritmo Prevh

0,6316 ou 63% de chance de X ser uma entidade da cor verde.
0,7369 ou 73% de chance de X ser uma entidade da cor vermelha.
0,6316 ou 63% de chance de X ser uma entidade da cor azul.

4.3. Conclusão dos testes

É possível perceber que enquanto o algoritmo KNN, foi inconclusivo para $K = 3$, o algoritmo Prevh conseguiu propor uma classificação vermelha a entidade X. Este fator ocorreu devido a diferente conotação que é dada a distância euclidiana em ambos os algoritmos, pois por mais que a relevância das informações presentes em $K = 3$ seja a mesma, o algoritmo Prevh utilizou-se diretamente da distância euclidiana no cálculo do índice de pertinência, enquanto o algoritmo KNN utilizou-se da distância euclidiana somente para definir o espaço K limitante.

5. Considerações Finais

As aplicações do algoritmo Prevh mais visíveis são em situações semelhantes as aplicações do algoritmo KNN, tanto quanto em situações que envolvam dados discretos definidos por processos estocásticos [3], pois, a relevância atribuída ao dado pode contribuir para a sequencialidade dos dados no tempo possibilitando a previsão de diversas direções para os quais os dados podem evoluir.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Referências

[1] A. Soofi, A. Awan, Classification Techniques in Machine Learning: Applications and Issues, Journal of Basic & Applied Sciences 13 (2017) 459 – 465. URL: https://pdfs.semanticscholar.org/2678/e213cec548d278879ceaf01582ee8913cc3f.pdf.

[2] B. Marr, M. Ward, Artificial Intelligence in Practice: How 50 Successful Companies Used AI and Machine Learning to Solve Problems, Wiley, 2019.

[3] M. KAC, J. LOGAN, Chapter 1 - fluctuations based in part on a series of lectures given at l'ecole polytechnique federale at lausanne, switzerland as a part of their troisieme cycle, summer 1974., em: E. MONTROLL, J. LEBOWITZ (Eds.), Fluctuation Phenomena, Elsevier, 1979, pg. 1–60.

[4] I. Newton, J. Rodrigues, F. C. Gulbenkian, S. de Educação e Bolsas, Princípios matemáticos da filosofia natural, Fundação Calouste Gulbenkian Serviço de Educação e Bolsas, Lisboa, 2010. OCLC: 959175701.

[5] J. Bezerra, Teorema de pitagoras no espaço, Revista do Professor de Matemática 79 (2012) 22–24.

[6] T. M. Mitchell, Machine learning, McGraw-Hill series in Computer Science, mcgraw-hill science/engineering/math; 1ª edição (1 março 1997) ed., McGraw-Hill, Porland, 1997.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Imputação de dados faltantes em base de dados de classificação hierárquica multirrótulo usando regressão linear e polinomial

Alvaro Mateus Santana[a], Helyane Bronoski Borges[a]

[a]*Universidade Tecnológica Federal do Paraná, Universidade de Tecnologia, Rua Doutor Washington Subtil Chueire, 330, Jardim Carvalho, Ponta Grossa, PR, Brasil*
alvarosantana@alunos.utfpr.edu.br

Resumo

Algoritmos de aprendizagem de máquina normalmente necessitam de um conjunto sem dados faltantes, fato que nem sempre ocorre. Considerando isso, se torna necessário usar métodos para imputar dados em conjuntos de dados, tornando, em muitas vezes, o processo de aprendizagem de dados mais eficiente. Este trabalho propõe a imputação de dados faltantes no cenário de classificação hierárquica multirrótulo, em base de dados do projeto Gene Ontology, utilizando regressão linear e regressão polinomial. Por meio dos experimentos foi possível observar que houve melhora na métrica AUPRC em uma das bases de dados testadas.

Abstract

Machine learning algorithms usually require a set with no missing data, which is not always the case. Considering this, it becomes necessary to use methods to impute data into datasets, often making the data learning process more efficient. This work proposes the imputation of missing data in the multi-label hierarchical classification scenario, in a Gene Ontology project database, using linear regression and polynomial regression. Through the experiments it was possible to observe that there was an improvement in the AUPRC metric in one of the tested databases.

Palavras-chaves: dados faltantes, classificação hierárquica multirrótulo, regressão

1. Introdução

Segundo Cerri [1], um dos problemas enfrentados pelas tarefas de classificação é quando os dados apresentam atributos faltantes. Técnicas de aprendizagem de máquina, como por exemplo árvore de decisão, podem gerar erro de execução quando um ou mais atributos do conjunto de treinamento não apresentam valor.

Para Farhangfar et al. [2], como os dados podem ser gerados de formas variadas, muitas vezes são

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

apresentadas bases com informações incompletas ou faltantes, ou seja, boa parte dos bancos de dados existentes é caracterizada pela imprecisão e pela incompletude, isto é, pela presença de valores ruidosos e faltantes, respectivamente.

Neste trabalho é proposto um método de imputação de dados faltantes em bases de dados com classificação hierárquica multirrótulo, utilizando regressão linear e polinomial.

2. Referencial teórico

Dados faltantes

Conjuntos de dados podem apresentar dificuldades relacionadas à qualidade, exemplos mais frequentes dessas dificuldades são dados ruidosos, inconsistentes, redundantes ou incompletos, a Figura 1 ilustra um conjunto de dados onde alguns atributos apresentam dados faltantes, sendo esses representados pelo caractere "?". No exemplo, caso fosse preciso aplicar determinado algoritmo de aprendizagem, como por exemplo Rede Neural Artificial, será necessário realizar alguma abordagem para tratar estes dados faltantes.

Id.	Atrib 1	Atrib 2	Atrib 3	Classe
1	Sim	Grande	125	Não
2	?	Médio	100	Não
3	Não	Pequeno	70	Não
4	Sim	Médio	?	Não
5	Não	Grande	95	Sim
6	Não	?	60	Não
7	Sim	Grande	?	Não
8	Não	Pequeno	85	Sim
9	Não	Médio	75	Não
10	Sim	Pequeno	90	Sim

Figura 1: Conjunto com dados faltantes. Tan et al. [3]

Classificação hierárquica multirrótulo

Para Cerri [1], existe um grande número de problemas em que uma ou mais classes podem ser divididas em subclasses ou agrupadas em superclasses, sendo um problema de classificação hierárquica. Segundo Borges [4], a hierarquia das classes pode ser representada por uma árvore ou DAG (*Direct Acyclic Graph*), sendo que a segunda abordagem se difere da primeira devido a ser possível uma classe possuir mais de um pai.

Carvalho et al. [5] diz que, normalmente na tarefa de classificação, cada instância é atribuída a somente um rótulo ou classe. Porém, em determinadas situações é necessário que dois ou mais rótulos sejam atribuídos para um exemplo, este tipo de classificação é denominado multirrótulo.

Por fim há determinados tipos de problemas que necessitam de uma representação utilizando ambas

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

as abordagens, sendo então um problema de Classificação Hierárquica Multirrótulo.

Regressão

Segundo Tan et al. [3], a regressão é uma técnica preditiva de análise de dados em que a variável alvo a ser avaliada é continua. Neste método, dado uma variável (ou um conjunto de variáveis) explicativas é possível estimar a variável alvo. A regressão tenta incorporar o conhecimento de outras variáveis, buscando produzir predições mais inteligentes.

Este trabalho propõe solucionar o problema de imputação utilizando um modelo de regressão linear ou polinomial.

Além dos conceitos de regressão é importante destacar o conceito de correlação que será de grande importância dentro do estudo proposto. O coeficiente de correlação mostra a relação entre as variáveis dependente e independente, sendo atribuído valor entre 0 e 1, sendo que quanto mais próximo de 1, maior será a correlação.

Base de dados

Para realização dos experimentos foram utilizados duas bases de dados do projeto *Gene Ontology* [6], elas constam nos trabalhos de Vens et al [7] e Borges [4]. O projeto possui três categorias de funções denominadas de: processo biológico, função molecular e componente celular que são implementadas em três ontologias independentes. Nessas ontologias, genes e proteínas podem ser classificadas de forma hierárquica, podendo apresentar mais de uma função ou característica. Nesta área a principal motivação do uso da classificação hierárquica multirrótulo é a possibilidade de realizar diagnóstico de doenças e desenvolvimento de medicamentos.

A Tabela 1 lista as duas bases de dados utilizadas, apresentando também características de cada.

Tabela 1 - Características das bases utilizadas nos experimentos preliminares

ID	Nome	Amostras	Atributos	Classes	Atributos com dados faltantes	Amostras com dados faltantes
1	Cellcycle	3751	77	4125	16136	3507
3	Eisen	2418	79	3573	3686	1823

O estado da arte

Foi realizada revisão da literatura acerca de métodos de imputação, porém não foi possível verificar abordagens desenvolvidas exclusivamente para o cenário de classificação hierárquica multirrótulo. Como alternativa, em um segundo eixo de revisão foi inspecionado como foi realizada a etapa de

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

preparação dos dados nos trabalhos relacionados a classificadores hierárquicos multirrótulo, sendo que neste eixo foram encontradas menções ao uso da média como método de imputação.

3. Método proposto e resultados dos experimentos realizados

A figura 2 ilustra as três etapas principais do método de imputação proposto. Inicialmente é criado um subconjunto de dados com outros exemplos que compartilham classes com o exemplo com dado faltante. Nesta fase, caso não seja encontrado outro exemplo que compartilhe a classe com dado faltante, também é verificado classes ascendentes dentro da hierarquia.

Na segunda etapa é realizada a busca do atributo com melhor coeficiente de correlação com o atributo com dado faltante. Caso o coeficiente de correlação seja pequeno, a regressão de dados será substituída pela média. Será utilizada a definição de Cohen [8] para baixas correlações, onde serão aceitos somente valores acima de 0,3.

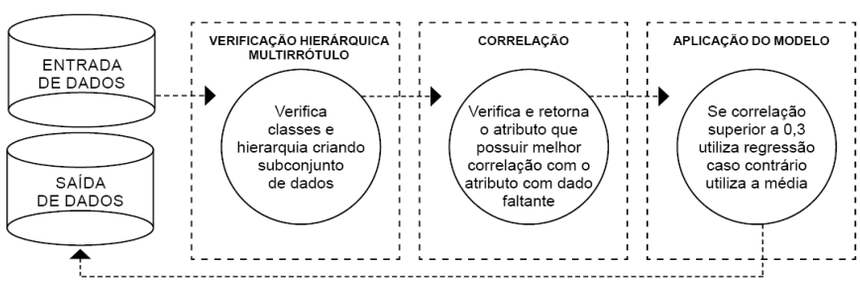


Figura 2: Fases do método de imputação

4. Experimentos e resultados

Os experimentos foram realizados em três etapas distintas e são ilustrados na Figura 3, inicialmente é utilizado a abordagem para tratar dados faltantes seguido do uso do classificador Clus-HMC proposto por Vens et al [7], na base de dados após a imputação de valores faltantes e por fim verificar a acurácia da classificação por meio da medida baseada na curva de precisão e revocação (AUPRC).

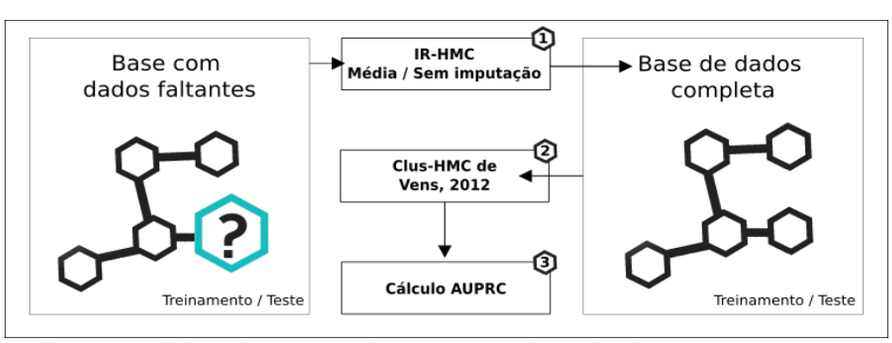


Figura 3: Metodologia utilizada na realização dos experimentos

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

Na execução do classificador Clus-HMC, as bases de dados foram normalizadas utilizando a técnica MinMax e divididas em 2/3 para o treinamento e 1/3 para o teste.

Os resultados foram comparados entre as duas abordagens do método (regressão linear e polinomial), com a média ponderada proposta por Borges [4] e também com a execução do algoritmo Clus-HMC sem imputação de dados faltantes. A Figura 4 ilustra os resultados obtidos.

	AUPRC - Média ponderada (Borges 2012)	AUPRC Sem imputação	AUPRC Regressão Linear	AUPRC Regressão Polinomial
Cellcycle	0,4375	0,4377	0,4394	0,4397
Eisen	0,4541	0,4570	0,4536	0,4537

Figura 4: Resultados dos experimentos

Observa-se que na base de dados Cellcycle a regressão linear e polinomial apresentou melhores resultados da métrica AUPRC quando comparados a média ponderada e quando não houve imputação, já na base de dados Eisen não foi possível observar esta melhora. Um possível motivo é o fato da base de dados Eisen possuir uma quantidade menor de amostras, o que pode ter afetado a acurácia do cálculo da correlação. Outro fato importante a ser analisado nos resultados é que a regressão foi utilizada 2.748 vezes nas amostras da base de dados Eisen que corresponde a 79% dos dados faltantes, enquanto na base Cellcycle o uso da regressão foi acionado 6.302 vezes que corresponde a 12% dos dados faltantes.

A avaliação dos resultados obtidos foi feita utilizando o teste de hipótese de Friedman [9], sendo que este teste é indicado para comparar o desempenho de vários algoritmos em diferentes bases de dados. O uso deste teste é indicado pois ele é não paramétrico, já que há dificuldade em se conhecer a distribuição dos dados. De acordo com os autores Iman e Davenport [10] o teste de Friedman é considerado muito conservador, e por isso, sugere o uso da estatística FF distribuída de acordo com a distribuição F de Snedecor.

Analisando os resultados apresentados na Figura 4, têm-se $F_x^2 = 1{,}2$. Usando a estatística F_F para correção de F_x^2 têm-se $F_F = 0{,}25$.

Para os graus de liberdade $k-1$ e $(k-1)*(n-1)$ em que $k=2$ e $n=4$, encontra-se o seguinte valor tabelado na distribuição de F de Snedecor $F(1{,}3) = 10{,}13$. Como $F_F < F(1{,}3)$, logo a hipótese nula não pode ser rejeitada.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

5. Conclusão

Com base nestes resultados, o método apresentou melhora na métrica AUPRC no conjunto de dados que possui quantidade maior de exemplos, além disso é importante observar o fato de a quantidade maior de exemplos também reduzir a quantidade de vezes que a regressão foi utilizada.

Porém, apesar de haver melhora na métrica AUPRC, o teste de Friedman mostrou que a hipótese nula não pode ser rejeitada, indicando que não há diferença estatística entre os resultados dos algoritmos.

Como trabalhos futuros podem ser realizados experimentos nas demais bases de dados do projeto Gene Ontology, além de serem testadas outras abordagens de regressão, como por exemplo regressão múltipla. Nestes experimentos a regressão polinomial foi testada com um polinômio quadrático, porém, podem ser realizados experimentos com polinômios cúbicos e superiores.

Referências

[1] R. Cerri. Técnicas de classificação hierárquica multirrótulo. 2010. 241 f. Dissertação (Mestrado em Ciência da Computação) - Instituto de Ciências Matemáticas e de Computação. Universidade de São Paulo, São Carlos, 2010.

[2] A. Farhangfar, L. A. Kurgan, W. Pedrycz. Experimental analysis of methods for imputation of missing values in databases. Proceedings of SPIE, Orlando, v. 5421, p. 172-182, 2004.

[3] P. Tan, M. Stenbach, V. Kumar, Introdução ao Data Mining Mineração de Dados. Rio de Janeiro: Editora Ciência Moderna, 2009.

[4] H. B. Borges, Classificador hierárquico multirrótulo usando uma rede neural competitiva. 2012. 188 f. Tese (Doutorado) - Universidade Católica do Paraná, Curitiba.

[5] A. C. P. L. F. Carvalho, J. Gama, A. C. Lorena, K. Faceli. Inteligência Artificial: uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC Editora, 2011.

[6] Gene Ontology, Gene Ontology Overview, disponível em http://geneontology.org/docs/ontology-documentation/. Acessado em 22/08/2021.

[7] C. Vens, J. Sruyf, L. Schietgat, S. Džeroski, Decision trees for hierarchical multi-label classification. Machine learning, Springer, v.73, n.2, p.185.

[8] J, Cohen, Statistical Power Analysis for the Behavioral Sciences. 2. ed. New York: Laurence Erlbaum Associates, 1988.

[9] J. H. Friedman, J. W. Tukey. A Projection Pursuit Algorithm for Exploratory Data Analysis. IEEE Transactions on Computers, IEEE, v. 100, n. 9, p. 881-890, 1974.

[10] R. L. Iman, J. M. Davenport. Approximations of the critical region of the Friedman statistic. Communications in Statistics. p. 571-595, 1980.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322

DOI:10.5281/zenodo.6090322

ISSN 2526-1371

Departamento Acadêmico de Informática
Universidade Tecnológica Federal do Paraná