

EDUCAÇÃO MATEMÁTICA E SUAS TECNOLOGIAS 2

Felipe Antonio Machado Fagundes Gonçalves
(Organizador)

 **Atena**
Editora
Ano 2019

Felipe Antonio Machado Fagundes Gonçalves
(Organizador)

Educação Matemática e suas Tecnologias 2

Atena Editora
2019

2019 by Atena Editora
Copyright © Atena Editora
Copyright do Texto © 2019 Os Autores
Copyright da Edição © 2019 Atena Editora
Editora Executiva: Prof^a Dr^a Antonella Carvalho de Oliveira
Diagramação: Natália Sandrini
Edição de Arte: Lorena Prestes
Revisão: Os Autores

O conteúdo dos artigos e seus dados em sua forma, correção e confiabilidade são de responsabilidade exclusiva dos autores. Permitido o download da obra e o compartilhamento desde que sejam atribuídos créditos aos autores, mas sem a possibilidade de alterá-la de nenhuma forma ou utilizá-la para fins comerciais.

Conselho Editorial

Ciências Humanas e Sociais Aplicadas

Prof. Dr. Álvaro Augusto de Borba Barreto – Universidade Federal de Pelotas
Prof. Dr. Antonio Carlos Frasson – Universidade Tecnológica Federal do Paraná
Prof. Dr. Antonio Isidro-Filho – Universidade de Brasília
Prof. Dr. Constantino Ribeiro de Oliveira Junior – Universidade Estadual de Ponta Grossa
Prof^a Dr^a Cristina Gaio – Universidade de Lisboa
Prof. Dr. Deyvison de Lima Oliveira – Universidade Federal de Rondônia
Prof. Dr. Gilmei Fleck – Universidade Estadual do Oeste do Paraná
Prof^a Dr^a Ivone Goulart Lopes – Istituto Internazionale delle Figlie de Maria Ausiliatrice
Prof^a Dr^a Juliane Sant’Ana Bento – Universidade Federal do Rio Grande do Sul
Prof. Dr. Julio Candido de Meirelles Junior – Universidade Federal Fluminense
Prof^a Dr^a Lina Maria Gonçalves – Universidade Federal do Tocantins
Prof^a Dr^a Natiéli Piovesan – Instituto Federal do Rio Grande do Norte
Prof^a Dr^a Paola Andressa Scortegagna – Universidade Estadual de Ponta Grossa
Prof. Dr. Urandi João Rodrigues Junior – Universidade Federal do Oeste do Pará
Prof^a Dr^a Vanessa Bordin Viera – Universidade Federal de Campina Grande
Prof. Dr. Willian Douglas Guilherme – Universidade Federal do Tocantins

Ciências Agrárias e Multidisciplinar

Prof. Dr. Alan Mario Zuffo – Universidade Federal de Mato Grosso do Sul
Prof. Dr. Alexandre Igor Azevedo Pereira – Instituto Federal Goiano
Prof^a Dr^a Daiane Garabeli Trojan – Universidade Norte do Paraná
Prof. Dr. Darllan Collins da Cunha e Silva – Universidade Estadual Paulista
Prof. Dr. Fábio Steiner – Universidade Estadual de Mato Grosso do Sul
Prof^a Dr^a Girlene Santos de Souza – Universidade Federal do Recôncavo da Bahia
Prof. Dr. Jorge González Aguilera – Universidade Federal de Mato Grosso do Sul
Prof. Dr. Ronilson Freitas de Souza – Universidade do Estado do Pará
Prof. Dr. Valdemar Antonio Paffaro Junior – Universidade Federal de Alfenas

Ciências Biológicas e da Saúde

Prof. Dr. Gianfábio Pimentel Franco – Universidade Federal de Santa Maria
Prof. Dr. Benedito Rodrigues da Silva Neto – Universidade Federal de Goiás
Prof.^a Dr.^a Elane Schwinden Prudêncio – Universidade Federal de Santa Catarina
Prof. Dr. José Max Barbosa de Oliveira Junior – Universidade Federal do Oeste do Pará
Prof.^a Dr.^a Natiéli Piovesan – Instituto Federal do Rio Grande do Norte
Prof.^a Dr.^a Raissa Rachel Salustriano da Silva Matos – Universidade Federal do Maranhão
Prof.^a Dr.^a Vanessa Lima Gonçalves – Universidade Estadual de Ponta Grossa
Prof.^a Dr.^a Vanessa Bordin Viera – Universidade Federal de Campina Grande

Ciências Exatas e da Terra e Engenharias

Prof. Dr. Adélio Alcino Sampaio Castro Machado – Universidade do Porto
Prof. Dr. Eloi Rufato Junior – Universidade Tecnológica Federal do Paraná
Prof. Dr. Fabrício Menezes Ramos – Instituto Federal do Pará
Prof.^a Dr.^a Natiéli Piovesan – Instituto Federal do Rio Grande do Norte
Prof. Dr. Takeshy Tachizawa – Faculdade de Campo Limpo Paulista

Conselho Técnico Científico

Prof. Msc. Abrãao Carvalho Nogueira – Universidade Federal do Espírito Santo
Prof.^a Dr.^a Andreza Lopes – Instituto de Pesquisa e Desenvolvimento Acadêmico
Prof. Msc. Carlos Antônio dos Santos – Universidade Federal Rural do Rio de Janeiro
Prof.^a Msc. Jaqueline Oliveira Rezende – Universidade Federal de Uberlândia
Prof. Msc. Leonardo Tullio – Universidade Estadual de Ponta Grossa
Prof. Dr. Welleson Feitosa Gazel – Universidade Paulista
Prof. Msc. André Flávio Gonçalves Silva – Universidade Federal do Maranhão
Prof.^a Msc. Renata Luciane Polsaque Young Blood – UniSecal
Prof. Msc. Daniel da Silva Miranda – Universidade Federal do Pará

Dados Internacionais de Catalogação na Publicação (CIP) (eDOC BRASIL, Belo Horizonte/MG)	
E24	Educação matemática e suas tecnologias 2 [recurso eletrônico] / Organizador Felipe Antonio Machado Fagundes Gonçalves. – Ponta Grossa (PR): Atena Editora, 2019. – (Educação Matemática e suas Tecnologias; v. 2) Formato: PDF Requisitos de sistema: Adobe Acrobat Reader Modo de acesso: World Wide Web Inclui bibliografia ISBN 978-85-7247-348-4 DOI 10.22533/at.ed.484192405 1. Matemática – Estudo e ensino – Inovações tecnológicas. 2. Tecnologia educacional. I. Gonçalves, Felipe Antonio Machado Fagundes. II. Série. CDD 510.7
Elaborado por Maurício Amormino Júnior – CRB6/2422	

Atena Editora
Ponta Grossa – Paraná - Brasil
www.atenaeditora.com.br
contato@atenaeditora.com.br

APRESENTAÇÃO

A obra “Educação Matemática e suas tecnologias” é composta por quatro volumes, que vêm contribuir de maneira muito significativa para o Ensino da Matemática, nos mais variados níveis de Ensino. Sendo assim uma referência de grande relevância para a área da Educação Matemática. Permeados de tecnologia, os artigos que compõem estes volumes, apontam para o enriquecimento da Matemática como um todo, pois atinge de maneira muito eficaz, estudantes da área e professores que buscam conhecimento e aperfeiçoamento. Pois, no decorrer dos capítulos podemos observar a matemática aplicada a diversas situações, servindo com exemplo de práticas muito bem sucedidas para docentes da área. A relevância da disciplina de Matemática no Ensino Básico e Superior é inquestionável, pois oferece a todo cidadão a capacidade de analisar, interpretar e inferir na sua comunidade, utilizando-se da Matemática como ferramenta para a resolução de problemas do seu cotidiano. Sem dúvidas, professores e pesquisadores da Educação Matemática, encontrarão aqui uma gama de trabalhos concebidos no espaço escolar, vislumbrando possibilidades de ensino e aprendizagem para diversos conteúdos matemáticos. Que estes quatro volumes possam despertar no leitor a busca pelo conhecimento Matemático. E aos professores e pesquisadores da Educação Matemática, desejo que esta obra possa fomentar a busca por ações práticas para o Ensino e Aprendizagem de Matemática.

Felipe Antonio Machado Fagundes Gonçalves

SUMÁRIO

CAPÍTULO 1	1
O ALGORITMO ESPECTRAL COMO ALTERNATIVA AO ALGORITMO K-MEANS EM CONJUNTO DE DADOS ARTIFICIAIS	
Luciano Garim Garcia Leonardo Ramos Emmendorfer	
DOI 10.22533/at.ed.4841924051	
CAPÍTULO 2	16
NOVAS RELAÇÕES NA MATRIZ DE TRANSFORMAÇÃO DA TRANSFORMADA NUMÉRICA DE PASCAL	
Arquimedes José De Araújo Paschoal Ricardo Menezes Campello De Souza Hélio Magalhães De Oliveira	
DOI 10.22533/at.ed.4841924052	
CAPÍTULO 3	24
ALGORITMOS RÁPIDOS PARA O CÁLCULO DA TRANSFORMADA NUMÉRICA DE PASCAL	
Arquimedes José De Araújo Paschoal Ricardo Menezes Campello De Souza	
DOI 10.22533/at.ed.4841924053	
CAPÍTULO 4	32
ANÁLISE DE CÁLCULO DIFERENCIAL USANDO O SOFTWARE GEOGEBRA	
Amanda Barretos Lima Garuth Brenda Anselmo Mendes Isabela Geraldo Reghin Rosângela Teixeira Guedes	
DOI 10.22533/at.ed.4841924054	
CAPÍTULO 5	46
DEFLEXÃO EM VIGAS DE CONCRETO ARMADO SOLUÇÃO ANALÍTICA E NUMÉRICA VIA MÉTODO DAS DIFERENÇAS FINITAS	
Mariana Coelho Portilho Bernardi Adilandri Mércio Lobeiro Jeferson Rafael Bueno Thiago José Sepulveda da Silva	
DOI 10.22533/at.ed.4841924055	
CAPÍTULO 6	57
MODELO MATEMÁTICO PARA AUXILIAR O PLANEJAMENTO DA MANUTENÇÃO PREVENTIVA DE MOTORES ELÉTRICOS	
Thalita Monteiro Obal Jonatas Santana Obal	
DOI 10.22533/at.ed.4841924056	

CAPÍTULO 7	64
PRINCÍPIO DA SUPERPOSIÇÃO E SOLUÇÃO NUMÉRICA DO PROBLEMA DE FLUXO EM AQUÍFERO CONFINADO	
João Paulo Martins dos Santos Alessandro Firmiano de Jesus Edson Wendland	
DOI 10.22533/at.ed.4841924057	
CAPÍTULO 8	83
RESONANT ORBITAL DYNAMICS OF CBERS SATELLITES	
Jarbas Cordeiro Sampaio Rodolpho Vilhena de Moraes Sandro da Silva Fernandes	
DOI 10.22533/at.ed.4841924058	
CAPÍTULO 9	91
TESTES ADAPTATIVOS ENVOLVENDO O CONTEÚDO DE DERIVADAS: UM ESTUDO DE CASO COM ALUNOS DE ENGENHARIA CIVIL	
Patrícia Liane Grudzinski da Silva Claudia Lisete Oliveira Groenwald	
DOI 10.22533/at.ed.4841924059	
CAPÍTULO 10	104
LOCALIZAÇÃO DE FALTAS EM LINHAS DE TRANSMISSÃO POR ANÁLISE DE SINAIS TRANSITÓRIOS DE TENSÃO	
Danilo Pinto Moreira de Souza Eliane da Silva Christo Aryfrance Rocha Almeida	
DOI 10.22533/at.ed.48419240510	
CAPÍTULO 11	116
MODELAGEM DA PROPAGAÇÃO DE FUMAGINA CAUSADA POR MOSCA-BRANCA EM CULTURAS AGRÍCOLA	
Gustavo Henrique Petrolí Norberto Anibal Maidana	
DOI 10.22533/at.ed.48419240511	
CAPÍTULO 12	133
LOS SUBNIVELES DE DESARROLLO DEL ESQUEMA DE DERIVADA: UN ESTUDIO EXPLORATORIO EN EL NIVEL UNIVERSITARIO	
Claudio Fuentealba Edelmira Badillo Gloria Sánchez-Matamoros Andrea Cárcamo	
DOI 10.22533/at.ed.48419240512	
CAPÍTULO 13	143
OTIMIZAÇÃO BASEADA EM CONFIABILIDADE PARA A MINIMIZAÇÃO DE FUNÇÕES MATEMÁTICAS	
Márcio Aurélio da Silva Fran Sérgio Lobato Aldemir Ap Cavalini Jr Valder Steffen Jr	
DOI 10.22533/at.ed.48419240513	

CAPÍTULO 14	156
SEQUÊNCIAS: INTERVALARES E FUZZY	
Gino Gustavo Maqui Huamán	
Ulcilea Alves Severino Leal	
Geraldo Nunes Silva	
DOI 10.22533/at.ed.48419240514	
CAPÍTULO 15	164
VALIDAÇÃO DO MÉTODO DOS ELEMENTOS DISCRETOS PARA O ESCOAMENTO DE GRÃOS DE SOJA	
Rodolfo França de Lima	
Vanessa Faoro	
Manuel Osório Binelo	
Dirceu Lima dos Santos	
Adriano Pilla Zeilmann	
DOI 10.22533/at.ed.48419240515	
CAPÍTULO 16	181
TAREAS DE GENERALIZACIÓN POR INDUCCIÓN PARA FORMAR EL CONCEPTO DE POTENCIA	
Landy Sosa Moguel	
Guadalupe Cabañas-Sánchez	
Eddie Aparicio Landa	
DOI 10.22533/at.ed.48419240516	
CAPÍTULO 17	192
SINCRONISMO EM UM NOVO MODELO METAPOPOPULACIONAL COM TAXA DE MIGRAÇÃO INDEPENDENTE DA DENSIDADE	
Francisco Helmuth Soares Dias	
Jacques Aveline Loureiro da Silva	
DOI 10.22533/at.ed.48419240517	
CAPÍTULO 18	199
SIMULAÇÃO 3D DO FLUXO DE AR DE UM SISTEMA REAL DE ARMAZENAGEM DE GRÃOS	
Vanessa Faoro	
Rodolfo França de Lima	
Aline Tampke Dombrowski	
Manuel Osório Binelo	
DOI 10.22533/at.ed.48419240518	
CAPÍTULO 19	207
CONTROLE ÓTIMO DO FLUXO DE ÁGUA EM UMA FÔRMA DE GELO	
Xie Jiayu	
João Luis Gonçalves	
DOI 10.22533/at.ed.48419240519	
CAPÍTULO 20	213
CÓDIGOS CÍCLICOS DEFINIDOS POR ANULAMENTO	
Conrado Jensen Teixeira	
Osnel Broche Cristo	
DOI 10.22533/at.ed.48419240520	

CAPÍTULO 21	216
ANÁLISE TEÓRICO-EXPERIMENTAL DE DISPERSÃO DE UM CONTAMINANTE COM TRANSFORMAÇÕES INTEGRAIS E INFERÊNCIA BAYESIANA	
Bruno Carlos Lugão Diego Campos Knupp Pedro Paulo Gomes Watts Rodrigues Antônio José da Silva Neto	
DOI 10.22533/at.ed.48419240521	
CAPÍTULO 22	225
ANÁLISE WAVELET DE TACOGRAMAS TEÓRICOS E EXPERIMENTAIS	
Ronaldo Mendes Evaristo Kelly Cristiane Iarosz Silvio Luiz Thomaz de Souza Ricardo Luiz Viana Moacir Fernandes de Godoy Antonio Marcos Batista	
DOI 10.22533/at.ed.48419240522	
CAPÍTULO 23	235
CONSTRUÇÃO DE UM AEROMODELO DE MACARRÃO NO ENSINO DE MATEMÁTICA E FÍSICA	
Alissan Sarturato Firão Ernandes Rocha de Oliveira Zulind Luzmarina Freitas	
DOI 10.22533/at.ed.48419240523	
SOBRE O ORGANIZADOR	239

O ALGORITMO ESPECTRAL COMO ALTERNATIVA AO ALGORITMO K-MEANS EM CONJUNTO DE DADOS ARTIFICIAIS

Luciano Garim Garcia

Universidade Federal do Rio Grande do Sul
(UFRGS)

Porto Alegre - Rio Grande do Sul

Leonardo Ramos Emmendorfer

Universidade Federal do Rio Grande (FURG)

Rio Grande - Rio Grande do Sul

RESUMO: Neste trabalho busca-se tratar do problema de agrupamento de dados pela abordagem de particionamento de grafos. Com isto, utiliza-se uma versão espectral e obtém-se uma solução relaxada que é utilizada no processo de agrupamento. No intuito de medir a eficácia do método espectral utiliza-se o algoritmo k-means para comparar resultados de agrupamento de dados artificiais. Finalmente, utiliza-se a medida-F para fazer o comparativo estatístico dos métodos e inferir conclusões.

PALAVRAS CHAVE: Particionamento, Agrupamento, Grafos, K-means.

ABSTRACT: In this work data clustering problem is transformed in graph partitioning problem. That way, a spectral version is used and a relaxed solution is obtained, and then is used in the clustering process. In order to measure the efficacy of the spectral method, the k-means algorithm is used to compare clustering results with artificial data. Finally, the F-measure is

used to make the statistical comparison of the methods and infer conclusions.

KEYWORDS: Partitioning, Clustering, Graphs, K-means.

1 | INTRODUÇÃO

O reconhecimento de padrões em matemática sempre foi alvo de estudo de vários pesquisadores. Nas últimas décadas diversos métodos de exploração de dados foram desenvolvidos com a finalidade de extrair informações úteis que antes estavam implícitas em números e tabelas. Sendo assim, pode-se dizer que estudo de agrupamento torna-se uma ferramenta útil para a análise de dados em diferentes situações. A análise de agrupamento tem aplicações em situações simples do dia-a-dia (desde o início da infância, aprende-se a distinguir entre gatos e cachorros, ou plantas e animais, melhorando continuamente esquemas de classificação subconscientes) assim como em diversas áreas, como biologia, pesquisa de mercado, processamento de imagens, reconhecimento de padrões, geografia, e muitas outras (Dos Santos, 2015).

Diversos métodos de agrupamento de dados buscam encontrar agrupamentos ótimos, porém surgem certas limitações correspondentes a cada algoritmo utilizado.

Com isto, existe uma grande necessidade de encontrar métodos inteligentes que realizem esta tarefa de maneira bem geral, independente do formato do conjunto de dados.

Neste trabalho, utiliza-se o método de agrupamento espectral baseado em particionamento grafos para descobrir os grupos existentes em conjuntos de dados artificiais. Desse modo, espera-se que os agrupamentos aconteçam de forma mais natural do que outros algoritmos como, por exemplo, o k-means (LÄNGKVIST, 2014).

2 | METODOLOGIA

Considere um conjunto de dados descritos por pares ordenados $(x_i, x_j) \in \mathbb{R}^2$. Cada par de elementos descrevem um ponto no plano cartesiano e, sua relação com os demais pode ser descrita por uma função de similaridade. Seja $W_{ij} = f(d(x_i, x_j), \sigma)$ a matriz simétrica composta pela função de similaridade que leva em consideração a distância euclidiana entre os elementos e parâmetro de distribuição gaussiana dos dados (NG, 2002). Assim pode-se modelar a matriz de similaridade de acordo com a equação abaixo:

$$W_{ij} = e^{-\frac{d^2(x_i, x_j)}{2\sigma^2}}$$

Neste caso, usando a matriz de similaridade dos dados poder-se-ia aplicar algum algoritmo que tomasse as informações dessa matriz para determinar os grupos. Porém, esta formulação é bastante sensível ao parâmetro que é uma medida global, não levando em consideração informações locais dos dados. Nesse sentido propõe-se transformar o problema de agrupamento de dados em um problema de particionamento de dados, conforme descrito na Figura 1.

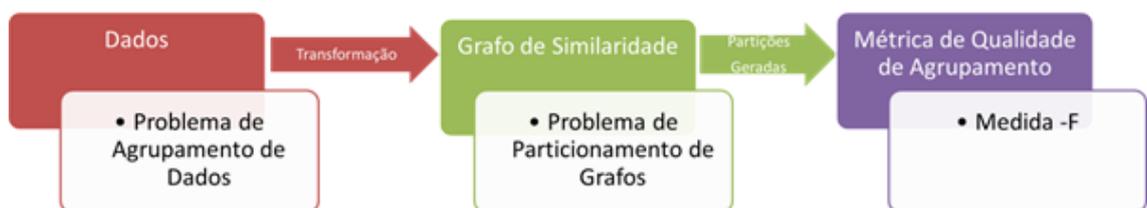


Figura 1- Esquema da metodologia abordada

Para fazer a transformação do problema considera-se cada par ordenado como o vértice do grafo e a medida de similaridade entre os dados é posto como o peso de cada aresta que conecta um vértice em outro. Neste sentido a nova formulação trata de obter partições no grafo que descrevam coerentemente a estrutura dos grupos a serem formados. No ponto de vista de particionamento utilizam-se funções

objetivo com o intuito de medir a qualidade das partições. Otimizada a função de particionamento e gerado os grupos, retorna-se ao problema de agrupamento com a solução obtida e, utiliza-se uma métrica de qualidade de agrupamento para medir a eficácia do resultado.

2.1 Solução do Problema de Particionamento de Grafos

Dado um número $k \in \mathbb{N}_{>1}$ e um grafo $G = (V, E)$ com pesos não negativos em suas arestas, $\omega: E \rightarrow \mathbb{R}_{>0}$, o problema de particionamento de grafos busca determinar uma partição P de V com grupos de vértices $P = \{V_1, \dots, V_k\}$ satisfazendo as seguintes condições: $V_1 \cup \dots \cup V_k = V$ e $V_i \cap V_j = \emptyset, \forall i \neq j$.

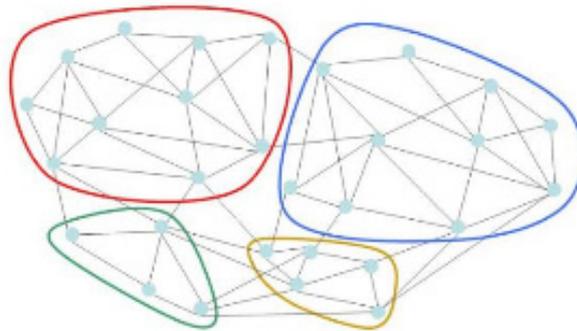


Figura 2- Grafo particionado

Considere a seguinte modelagem para o caso particular do biparticionamento: dados dois conjuntos disjuntos de vértices V e V^c de um grafo G define-se:

- O somatório dos pesos das arestas dos dois conjuntos por:

$$Cut(V, V^c) = \sum_{u \in V, v \in V^c} w_{uv}$$

- O somatório dos pesos das arestas dentro do conjunto V por:

$$Cut(V, V) = \sum_{u \in V, v \in V} w_{uv}$$

- O total de pesos provenientes do grupo V :

$$Vol(V) = \sum_{u \in V} d_u$$

Para medir a qualidade das partições é possível recorrer ao uso de funções objetivo que otimizadas propiciem partições equilibradas. Uma partição equilibrada ocorre quando vértices da mesma partição possuem arestas com pesos grandes (alta similaridade) e aresta entre partições distintas possuam pesos pequenos (pouca similaridade). Neste sentido, podem-se utilizar funções de corte mínimo, corte por

razão, e corte normalizado (JIA, 2014). Neste trabalho é utilizado a função de corte normalizado justamente por proporcionar cortes mais equilibrados no grafo (SHI, 2000).

2.1.1 Função de Corte Normalizado

O particionamento leva em consideração as conexões dentro e fora de um grupo, ou seja,

$$NCut(V, V^c) = Cut(V, V^c) \left(\frac{1}{Vol(V)} + \frac{1}{Vol(V^c)} \right)$$

Neste caso é levado em consideração o tamanho das partições. Porém, quando se introduz condições para obter partições equilibradas, o problema do corte mínimo embutido na função de corte normalizado torna-se um problema NP-difícil (WAGNER, 1993). Nesse sentido, trabalha-se com a versão relaxada do problema obtendo uma solução aproximada para a função objetivo.

Considere a seguinte relaxação para $NCut$:

Seja $f = [f_1, \dots, f_n]^t$ com $f_u = \begin{cases} \frac{1}{Vol(V)}, se u \in V \\ -\frac{1}{Vol(V^c)}, se u \in V^c \end{cases}$. A fim de reescrever $NCut$ considere:

$$f^t L f = \sum_{uv} w_{uv} (f_u - f_v)^2 = \sum_{u \in V, v \in V} w_{uv} \left(\frac{1}{Vol(V)} + \frac{1}{Vol(V^c)} \right)^2$$

Onde L é chamada de matriz laplaciana, definida por $L=D-W$, tal que W é a matriz de pesos do grafo (ou adjacências) e D é a matriz diagonal com elementos descritos por $Vol(V, V^c)$

Procedendo analogamente para a matriz D :

$$f^t D f = \sum_{u \in V} d_u f_u^2 = \sum_{u \in V} \frac{d_u}{Vol(V)^2} + \sum_{v \in V^c} \frac{d_u}{Vol(V^c)^2} = \frac{1}{Vol(V)} + \frac{1}{Vol(V^c)}$$

Então, reescreve-se $NCut$, como:

$$NCut(V, V^c) = \frac{f^t L f}{f^t D f}$$

Em resumo, tem-se de otimizar a seguinte função:

$$\left\{ \begin{array}{l} \min NCut(V, V^c) = \min \frac{f^t L f}{f^t D f}, \text{ onde} \\ f = [f_1, \dots, f_n]^t \text{ com } f_u = \begin{cases} \frac{1}{Vol(V)}, se u \in V \\ -\frac{1}{Vol(V^c)}, se u \in V^c \end{cases} \end{array} \right\}$$

Para esse propósito, considera-se: $\frac{f^t L f}{f^t D f}$ sujeito a $f^t D 1$ com solução dada pelo segundo menor autovalor do problema (VON LUXBURG, 2007).

$$L f = \lambda D f$$

$$D D^{-1} L f = \lambda D f$$

$$D^{-1} L f = \lambda f$$

$$\mathcal{L} f = \lambda f$$

De acordo com (FIEDLER, 1973), o segundo menor autovetor associado ao segundo menor autovalor, que é solução do problema acima, é capaz de descrever uma partição no grafo determinada pelo sinal das suas componentes. Ou seja, é feito um mapeamento dessas componentes diretamente no grafo gerando as partições. Em suma, ao invés de particionar o grafo de acordo com as componentes dos vetores característicos da solução exata utiliza-se as componentes da solução relaxada conforme a figura abaixo.

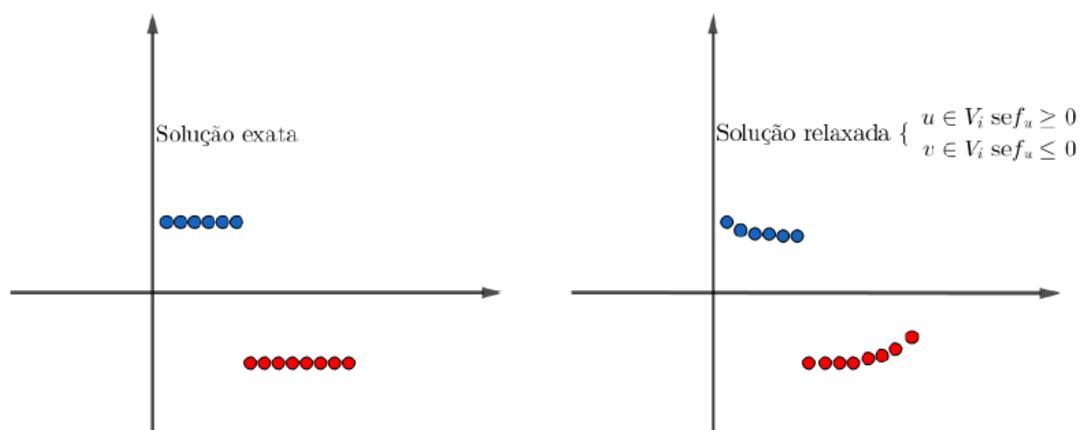


Figura 3 – Solução exata e solução relaxada

Dependendo do volume de dados (quantidade de vértices) é necessário aplicar um algoritmo de agrupamento nas componentes do autovetor solução para determinar os grupos. Em geral, utiliza-se o algoritmo k-means, amplamente utilizado na literatura, já que em poucas iterações é possível obter os grupos. Além disso, no caso geral de partições em subconjuntos disjuntos, utilizam-se as componentes dos k -menores autovetores para determinar as partições, sendo fundamental o uso de um algoritmo de agrupamento nesta etapa (NG, 2002).

3 | ALGORITMOS DE AGRUPAMENTO

Nas próximas seções são apresentados os algoritmos k-means e espectral via k-means, objetos de estudo deste trabalho.

3.1 Algoritmo K-means

O algoritmo k-means baseia-se na minimização de uma medida de custo, a distância interna entre os padrões de um agrupamento (LINDEN, 2009). A minimização do custo garante encontrar um mínimo local da função objetivo, que dependerá do ponto inicial do algoritmo. Esse tipo de algoritmo é chamado de ‘não-convexo’, pois, a cada iteração diminui o valor da distorção, visto que o resultado final depende do ponto inicial usado pelo algoritmo.

Algoritmo 1: Algoritmo k-means

Entrada: Conjunto de exemplos contendo vetores de atributos d-dimensionais e k= número de grupos.

Saída: k vetores de média e afiliação para os N vetores de atributos D

1. Escolha estimativas iniciais arbitrárias θ_j (0) para os θ_{jis} ; $j=1, \dots, m$ (isto é, para centroides dos k grupos);

2. Repita:

for $i \leftarrow 1$ to N do

 Determine o representante mais próximo, isto é, θ_j (centroide mais próximo) de x_i .

 Faça $b(i) = j$;

for $i \leftarrow 1$ to k do

 Atualização dos parâmetros: Determinar θ_j como a média dos vetores $x_i \in X$ com $b(i) = j$;

3. Repetir passo 2 até que não ocorram mudanças em θ_j entre duas iterações sucessivas.

A principal vantagem do k-means é a convergência da solução em poucas iterações, sendo esta convergência dependente da inicialização dos centros dos grupos e da geometria dos dados. Outro fator importante para o funcionamento do algoritmo é a informação preliminar da suposta quantidade de grupos contidos na distribuição dos dados. Ou seja, o número de k grupos deve também ser informado pelo usuário, acarretando uma dependência no resultado dos agrupamentos. Além disso, o k-means pode não funcionar de maneira esperada dependendo da curvatura da distribuição dos dados no plano bidimensional (GARCIA, 2017). Geralmente, o método funciona bem em conjuntos onde os grupos estão bem separados, ou seja, em distribuições gaussianas.

3.2 Algoritmo Espectral via K-means

O algoritmo pioneiro desta metodologia surgiu graças ao trabalho de (NG, 2002), que apresentou sua metodologia no artigo intitulado "On spectral clustering: Analysis

and an Algorithm”. Os passos do algoritmo espectral basicamente desempenham a tarefa de obter a matriz laplaciana do grafo de similaridade e calcular os seus autovetores para agrupá-los via k-means.

O parâmetro de escala σ controla a rapidez com que a matriz de similaridade decresce levando em consideração a distância entre x_i e x_j (VON LUXBURG, 2007). A matriz W é obtida calculando-se o peso de cada aresta incidente em um determinado vértice, desse modo, forma-se uma matriz diagonal, a qual é utilizada na obtenção da matriz laplaciana. Calculando-se os k autovetores de L , obtêm-se as matrizes U e logo após Y onde no passo 5 ocorre a aplicação do algoritmo k-means.

Algoritmo 2: Agrupamento espectral via k-means (NG, 2002)

Entrada: Conjunto de pontos $X = \{x_1, \dots, x_n\}$, desvio padrão σ e o número de grupos k .

Saída: Grupos A_1, \dots, A_k

1. Formar a matriz de similaridade W definida por: $W_{ij} = e^{-\frac{d^2(x_i, x_j)}{2\sigma^2}}$

2. Construir a matriz laplaciana: $L = D - W$

3. Encontrar os k autovetores de L (escolhidos para serem ortogonais entre si no caso de autovalores repetidos), e forme matriz U colocando os autovetores em colunas:

$$U[u_1 : \dots : u_k] \in \mathbb{R}^{n \times k}$$

4. Formar a matriz Y a partir de U normalizando cada linha de U para ter valores unitários.

5. Considerar cada linha de Y como um ponto em \mathbb{R}^k e classifique-os em k grupos via algoritmo k-means

$$Y_{ij} = \frac{u_{ij}}{[\sum_{j=1}^k u_{ij}^2]^{\frac{1}{2}}}$$

6. Colocar os pontos originais x_i no grupo j , se e somente se, a linha i da matriz Y for colocada no grupo j .

4 | EXPERIMENTOS

Nesta seção apresenta-se uma série de experimentos feitos com a aplicação dos dois algoritmos de agrupamentos em diversos conjuntos de dados. A utilização do k-means tem por objetivo obter uma comparação de resultados em relação ao algoritmo espectral, já tal algoritmo possui popularidade em diversas aplicações. Para medir a qualidade dos agrupamentos obtidos, optou-se pelo uso da medida-F, sendo que em todos os conjuntos utilizados é possível obter o seu "ground truth" na literatura.

4.1 Medida Estatística de Qualidade de Agrupamento

Com o intuito de medir o desempenho dos algoritmos utilizados em cada conjunto de dados foi utilizada a medida-F. Dada uma predição $(h(x) = (h_1(x), \dots, h_m(x))) \in Y$ de rótulo binário $y = (y_1, \dots, y_m)$, a medida-F é definida a seguir:

$$F(y, h(x)) = \frac{(1 + \beta^2) \sum_{i=1}^m y_i h_i(x)}{\beta^2 \sum_{i=1}^m y_i + \sum_{i=1}^m h_i(x)} \in [0,1]$$

Esta medida essencialmente corresponde a uma média harmônica ponderada de precisão e exaustividade (DEMBCZYNSKI, 2013). Para utilizar esta medida foram considerados os resultados corretos dos agrupamentos encontrados na literatura. Para cada teste foram definidas as variáveis TrueCluster e PredictedCluster, ou seja, a informação dos grupos corretos e dos grupos formados pelos algoritmos respectivamente.

4.2 Conjunto de Experimentos I

Em cada experimento a seguir existe um determinado problema de agrupamento relacionado à sua estrutura. O primeiro conjunto chamado Two Spirals possui curvaturas acentuadas que dificultam a aplicação dos algoritmos.

No segundo conjunto de dados ClusterinCluster existem dois grupos formados por duas concentrações circulares de pontos, uma pequena e outra maior. Neste caso, o desafio está em agrupar corretamente um grupo que está dentro de outro. Tal dificuldade é evidente, pois os conjuntos não são linearmente separáveis, além de possuírem variâncias distintas.

Por fim, no conjunto Corners existem quatro concentrações de pontos distribuídos em forma de esquinas. Os grupos deste conjunto possuem a característica de serem construídos de maneira bem simétrica, enfatizando claramente a sua divisão. Igualmente espaçados, os retângulos que formam as esquinas possuem a mesma distância em relação aos retângulos dos outros grupos adjacentes. Devido a esta característica, a dificuldade em agrupar as esquinas corretamente consiste em não levar apenas em consideração o fator distância entre grupos, mas também o fator vizinhança de cada elemento.

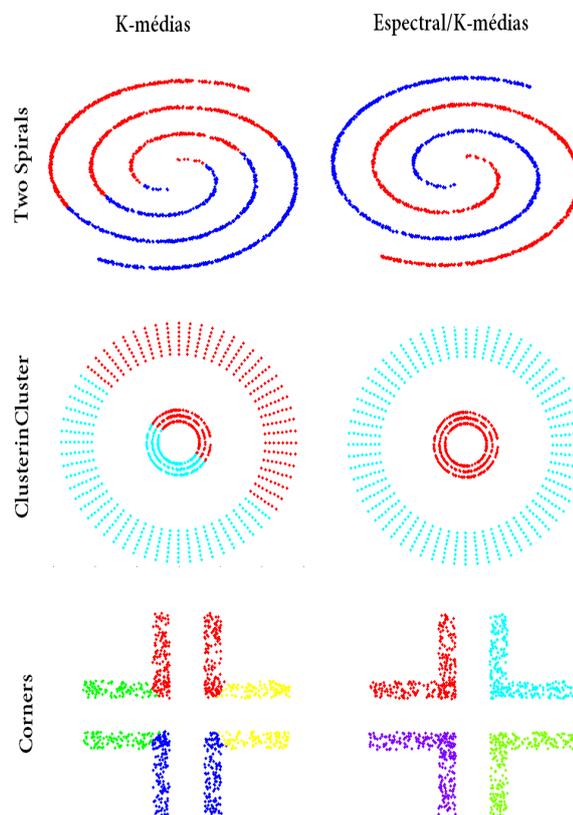


Figura 4 - Resultados dos experimentos referentes aos conjuntos de dados Two Spirals, ClusterinCluster e Corners para os algoritmos k-means e espectral.

A partir da Figura 4 pode-se obter uma visão geral da aplicação dos algoritmos nos conjuntos de dados em estudo. O experimento referente ao conjunto de dados Two Spirals mostrou a eficácia do método espectral em determinar corretamente os grupos. Já a aplicação do k-means resultou em uma partição linear no grafo, gerando também dois grupos, porém com resultados não satisfatórios.

No experimento referente ao conjunto ClusterinCluster, os resultados obtidos pelo k-means foram similares aos resultados do conjunto anterior. Uma partição linear descreve os dois grupos gerados de acordo com os centroides inicializados e posicionados iterativamente entre os dois grupos reais.

No último conjunto de dados Corners o método espectral via k-means obteve corretamente os grupos. Devido a simetria dos dados, o algoritmo k-means agrupou os retângulos de acordo com os centroides posicionados ao norte, sul, leste e oeste do conjunto, não retratando corretamente a estrutura dos grupos.

4.3 Conjunto de Experimentos II

Para a aplicação dos métodos de agrupamento discutidos aqui, foram escolhidos alguns conjuntos de dados disponíveis em (FU, 2007), (JAIN, 2005) e (CHANG, 2008). Houve a preocupação em trabalhar com conjuntos de dados de diferentes formatos geométricos, sendo específicos para a finalidade de analisar os resultados de cada algoritmo aplicado.

No primeiro teste foi utilizado o conjunto de dados Jain que possui o formato de duas meias luas conforme a Figura 5. Pode-se inferir visualmente que o conjunto possui dois grupos bem definidos e bem separados no plano bidimensional. A principal dificuldade em agrupar corretamente estes conjuntos está relacionada diretamente à curvatura em que os mesmos se dispõem no plano. Sendo assim, é interessante ressaltar que metodologias que obtenham um particionamento linear no conjunto provavelmente poderão falhar ao tentar descobrir corretamente os grupos.

No segundo experimento foi utilizado o conjunto de dados Flame, que por sua vez possui características geométricas interessantes, composta por uma distribuição esférica de elementos e outra concentração no formato de meia lua conforme Figura 5.

Por fim, é utilizado o conjunto de dados Spiral que também possui um formato de curva, neste caso é possível identificar que o número de grupos existentes no conjunto é três, conforme é possível analisar na Figura 5.

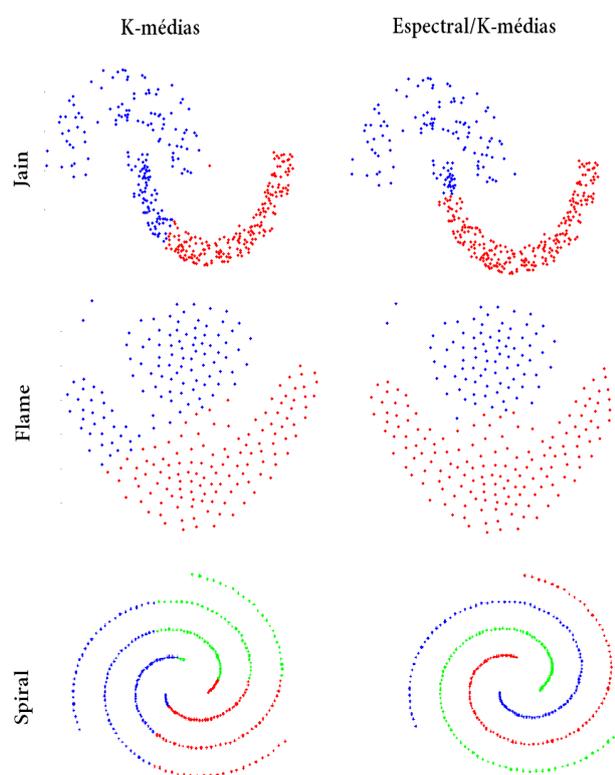


Figura 5 - Resultados dos experimentos referentes aos conjuntos de dados Jain, Flame e Spiral para os algoritmos k-means e espectral.

Conforme a visualização da Figura 5 é possível notar que o algoritmo k-means não obteve um agrupamento coerente de acordo com a estrutura do conjunto Jain. Isto se deve ao fato de que a utilização de centroides como parâmetro de agrupamento não funciona de maneira eficaz em conjuntos de dados que possuem uma determinada curvatura. Como a distribuição dos dados no plano não acompanha uma distribuição gaussiana, então o uso do k-means neste caso não é o mais recomendado. Já os grupos formados pelo algoritmo espectral geraram um resultado melhor que o seu antecessor. Apesar disso, a informação proveniente da matriz laplaciana não foi

suficiente para estabelecer um agrupamento coerente com a geometria do conjunto. Neste caso, existem duas possibilidades para a falha, a primeira seria a ineficácia do k-means quando aplicado nas componentes do autovetor solução e, a segunda a proximidade das componentes que podem levar a um agrupamento equivocado.

Os grupos do conjunto Flame estabelecidos pela metodologia k-means, formaram-se de maneira incorreta, agrupando parte da meia lua juntamente com a concentração superior acima dela. Considerando a concentração circular de elementos neste conjunto, conclui-se que o k-means consegue eficientemente colocar um centroide correto nesta região. Porém, em relação à meia lua este acerto do método torna-se impossível. No segundo algoritmo o resultado coerente, pois visualmente os grupos formaram-se corretamente respeitando a distribuição dos elementos no plano. É interessante ressaltar que o desvio padrão dos dados é moderadamente diferente entre os dois grupos, o que pode dificultar a resposta de vários algoritmos baseados neste parâmetro.

Novamente o número de grupos informado como entrada em ambos os algoritmos é o mesmo, desse modo, será possível estabelecer uma relação entre os resultados obtidos por cada um deles. No experimento com o uso do k-means o agrupamento do conjunto Spiral não foi satisfatório, novamente o problema ocorre devido à curvatura da distribuição dos elementos no plano. Diferente de muitos casos, neste conjunto nem uma boa inicialização do centro irá estabelecer corretamente os grupos, sendo assim, o algoritmo é ineficaz neste caso. Em contrapartida, o algoritmo baseado na metodologia de particionamento espectral apresentou bons resultados.

4.4 Conjunto de Experimentos III

Nesta seção de experimentos os conjuntos de dados foram escolhidos de acordo com a sua dimensão. Em experimentos anteriores os conjuntos utilizados eram bidimensionais, porém, propõe-se nesta seção a aplicação dos algoritmos em elementos com coordenadas (x_i, x_j, x_k) . A escolha dos dois primeiros bancos de dados foi de acordo com o "The Fundamental Clustering Problems Suite (FCPS)". Cada conjunto possui uma determinada característica que dificulta o processo de agrupamento. Problemas relacionados a densidade, distância, variância e não linearidade são abordados. No último experimento desta seção, é utilizado o conjunto de dados Iris, proveniente do repositório de Aprendizagem de Máquina UCI.

O conjunto de dados Chainlink é composto por dois anéis tridimensionais, tal que sua principal característica de agrupamento é a separação não linear dos grupos. Devido à proximidade entre os elementos dos dois grupos o fator distância não se torna suficiente para a determinação correta do agrupamento.

Analisando a Figura 6 nota-se, primeiramente, que o resultado de agrupamento pelo método k-means não obtém corretamente os grupos formados pelos dois anéis. Esta configuração do conjunto não permite que as posições dos centroides se ajustem

de acordo com a curvatura dos dados, desse modo, ocorre uma partição linear do conjunto gerando dois grupos que não correspondem à geometria correta dos dados.

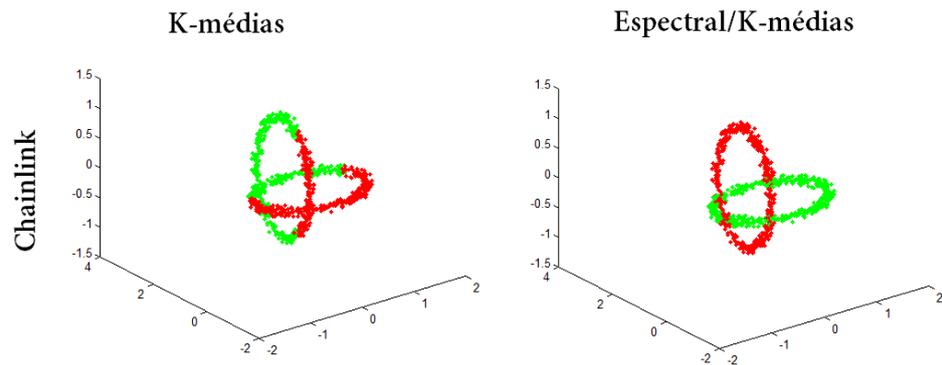


Figura 6 - Resultados dos experimentos referentes ao conjunto de dados Chainlink para os algoritmos k-means e espectral.

Nesse experimento o algoritmo espectral conseguiu identificar corretamente os grupos, isso se deve ao fato de que as componentes do autovetor solução estavam bem separadas, assim não houve dificuldades para o agrupamento e mapeamento dos mesmos.

Um problema interessante muito abordado em técnicas de agrupamento está relacionado a diferença de desvio-padrão e variância entre diferentes grupos. No caso do conjunto Atom, na Figura 7, existem dois grupos que possuem valores bem distintos de desvio-padrão, além de que existe o fato de um grupo menor estar contido dentro de outro maior. A característica do conjunto também não ser linearmente separável torna-se um agravante para o agrupamento correto dos elementos.

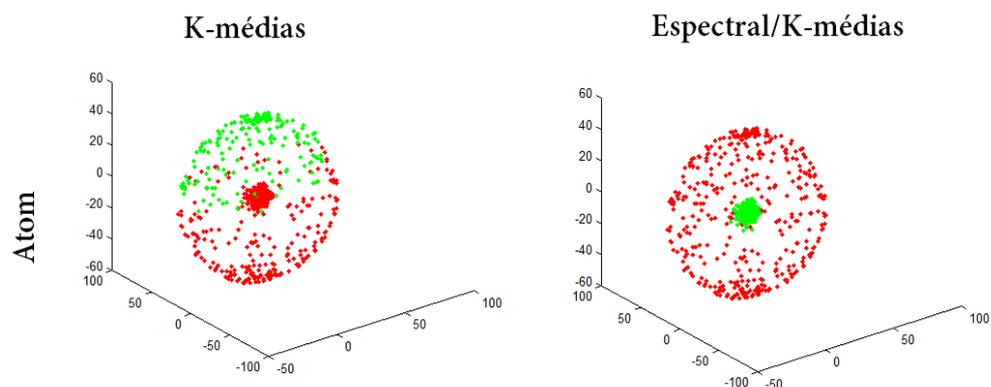


Figura 7 - Resultados dos experimentos referentes ao conjunto de dados Atom para os algoritmos k-means e espectral.

Novamente, o fator agrupamento via centroides não foi eficaz, evidenciando a dificuldade de estabelecer grupos em formato esférico também em dimensão

tridimensional. Em contrapartida, o algoritmo espectral obteve o resultado ótimo mesmo existindo a enorme diferença de desvio-padrão entre os grupos.

O conjunto Iris é um dos mais conhecidos nas áreas de classificação e agrupamento de dados. Introduzido por Ronald Fisher em 1936 no artigo "The use of multiple measurements in taxonomic problems" o seu uso traz importantes avanços na área de análise discriminante linear. O conjunto de dados consiste em 50 amostras de cada uma das três espécies de Iris (Iris setosa, Iris virginica e Iris versicolor). Foram medidas quatro características de cada amostra: o comprimento e a largura das sépalas e pétalas, em centímetros. Com base na combinação dessas quatro características, Fisher desenvolveu um modelo discriminante linear para distinguir as espécies umas das outras.

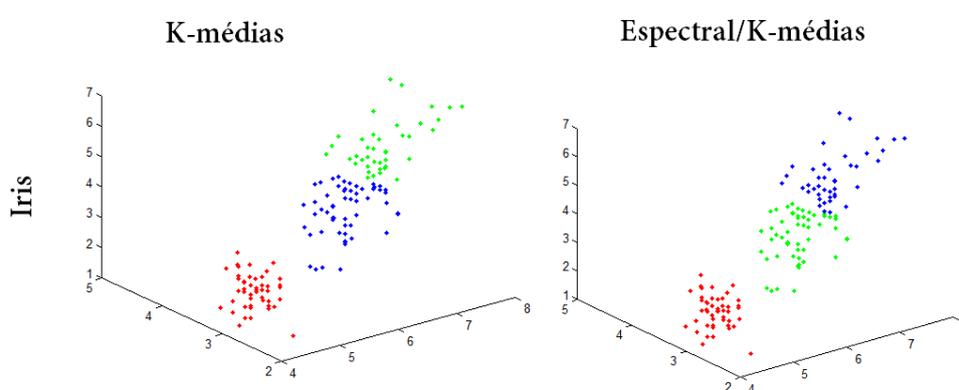


Figura 8 - Resultados dos experimentos referentes ao conjunto de dados Iris para os algoritmos k-means e espectral.

Na Figura 8 nota-se que os algoritmos k-means e espectral via k-medias apresentam resultados bem similares. Pode-se notar que grupo mais abaixo do conjunto é o mesmo em ambos os resultados de agrupamento dos algoritmos, neste caso, este grupo refere-se a Iris setosa. Desse modo, os dois métodos diferem apenas na parte superior do conjunto Iris, onde se concentram os grupos Iris virginica e a Iris versicolor, os quais apresentam uma maior concentração de pontos e proximidade entre si, dificultando a tarefa de agrupamento.

Conforme a tabela abaixo é possível obter um comparativo mais abrangente entre os dois algoritmos abordados neste trabalho. A tabela consta de resultados obtidos de medida-F quando utilizadas as saídas dos algoritmos, e comparadas com os grupos reais fornecidos pelos repositórios de dados utilizados. A partir dos valores da tabela pode-se dizer que o algoritmo k-means obteve uma média de 0,54 de medida-F, indo à contrapartida do algoritmo espectral com média de 0,97 de medida-F. Em geral, considerando conjuntos de dados com uma geometria mais simples, ambos algoritmos podem apresentar resultados similares, conforme pode-se ser visto no conjunto Iris. Porém, quando confrontados para aplicação em dados com certo grau de curvatura o algoritmo espectral é o que possui melhores resultados de medida-F.

Resultado da Medida-F		
Conjuntos de dados	K-means	Espectral
TwoSpirals	0,4797	1
ClusterinCluster	0,4138	1
Corners	0,2813	1
Jain	0,8288	0,9446
Flame	0,806	0,9825
Spiral	0,2695	1
Chainlink	0,384	1
Atom	0,6146	1
Iris	0,8085	0,8148

Tabela 1 – Resultados de medida-F para os conjuntos de dados utilizados nos experimentos.

5 | CONCLUSÃO

Neste trabalho foram apresentados conceitos introdutórios de particionamento espectral de grafos, juntamente com a apresentação do principal algoritmo de agrupamento espectral de dados. Pode-se notar a eficácia do método espectral quando utilizado em conjuntos de dados que possuam uma curvatura acentuada, enfatizando a sua utilização nesse contexto. Em diversos problemas de aplicação nem sempre os dados terão um formato viável para o seu agrupamento, sendo assim, o método espectral pode ser utilizado garantindo uma saída melhor do k-means proporcionaria.

No que se refere a medida de qualidade de agrupamento, nem sempre será possível utilizar a medida-F, visto que, em conjunto de dados reais não se sabe previamente a distribuição correta dos grupos. Seria interessante, para trabalhos futuros, estabelecer uma medida que permita ao algoritmo obter um feedback quanto a detecção do número correto de grupos reais. Algumas medidas já possuem tal propósito, como por exemplo, a medida Silhouette, que é amplamente utilizada nos algoritmos baseados no k-means.

Por fim, espera-se que este trabalho contribua como referencial conceitual sobre agrupamento de dados e possa ser utilizado para futuras contribuições nesta área de pesquisa. Além disso, enfatizando os bons resultados, conclui-se que a metodologia é promissora e pode ser aprimorada em trabalhos futuros.

REFERÊNCIAS

CHANG, Hong; YEUNG, Dit-Yan. **Robust path-based spectral clustering**. Pattern Recognition, v. 41, n. 1, p. 191-203, 2008.

DEMBCZYNSKI, Krzysztof et al. **Optimizing the F-measure in multi-label classification: Plug-in rule approach versus structured loss minimization**. In: International Conference on Machine Learning. 2013. p. 1130-1138.

DOS SANTOS, Daniel F.; DE MENDONÇA, Luziane F.; TEIXEIRA, Marcello G. **Uma algoritmo de agrupamento heterogêneo para formação de grupos de aprendizagem**. Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 3, n. 1, 2015.

FIEDLER, Miroslav. Algebraic connectivity of graphs. Czechoslovak mathematical journal, v. 23, n. 2, p. 298-305, 1973.

FU, Limin; MEDICO, Enzo. FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data. BMC bioinformatics, v. 8, n. 1, p. 3, 2007.

GARCIA, Luciano G.; EMMENDORFER, Leonardo R. Agrupamento Espectral Aglomerativo: Uma Proposta de Algoritmo. Anais do Computer on the Beach, p. 190-199, 2017.

JIA, Hongjie et al. The latest research progress on spectral clustering. Neural Computing and Applications, v. 24, n. 7-8, p. 1477-1486, 2014.

LÄNGKVIST, Martin; KARLSSON, Lars; LOUTFI, Amy. **A review of unsupervised feature learning and deep learning for time-series modeling**. Pattern Recognition Letters, v. 42, p. 11-24, 2014.

LINDEN, Ricardo. Técnicas de agrupamento. Revista de Sistemas de Informação da FSMA, v. 4, p. 18-36, 2009.

NG, Andrew Y.; JORDAN, Michael I.; WEISS, Yair. **On spectral clustering: Analysis and an algorithm**. In: Advances in neural information processing systems. 2002. p. 849-856.

SHI, Jianbo; MALIK, Jitendra. **Normalized cuts and image segmentation**. IEEE Transactions on pattern analysis and machine intelligence, v. 22, n. 8, p. 888-905, 2000.

VON LUXBURG, Ulrike. A tutorial on spectral clustering. Statistics and computing, v. 17, n. 4, p. 395-416, 2007.

WAGNER, Dorothea; WAGNER, Frank. **Between min cut and graph bisection**. In: International Symposium on Mathematical Foundations of Computer Science. Springer, Berlin, Heidelberg, 1993. p. 744-750.

SOBRE O ORGANIZADOR

FELIPE ANTONIO MACHADO FAGUNDES GONÇALVES Mestre em Ensino de Ciência e Tecnologia pela Universidade Tecnológica Federal do Paraná(UTFPR) em 2018. Licenciado em Matemática pela Universidade Estadual de Ponta Grossa (UEPG), em 2015 e especialista em Metodologia para o Ensino de Matemática pela Faculdade Educacional da Lapa (FAEL) em 2018. Atua como professor no Ensino Básico e Superior. Trabalha com temáticas relacionadas ao Ensino desenvolvendo pesquisas nas áreas da Matemática, Estatística e Interdisciplinaridade.

Agência Brasileira do ISBN
ISBN 978-85-7247-348-4

