UNIVERSIDADE FEDERAL DO RIO GRANDE
CENTRO DE CIÊNCIAS COMPUTACIONAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO
CURSO DE MESTRADO EM ENGENHARIA DE COMPUTAÇÃO

Dissertação de Mestrado

Agrupamento Espectral Aglomerativo: Uma Proposta de Algoritmo

Luciano Garim Garcia

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal do Rio Grande, como requisito parcial para a obtenção do grau de Mestre em Engenharia de Computação

Orientador: Prof. Dr. Leonardo Ramos Emmendorfer

Rio Grande, 2017

Ficha catalográfica

G216a Garcia, Luciano Garim.
Agrupamento espectral aglomerativo: uma proposta de algoritmo / Luciano Garim Garcia / Rosmer Haumán Vásquez. – 2017.
62 p.

Dissertação (mestrado) – Universidade Federal do Rio Grande – FURG, Programa de Pós-graduação em Engenharia de Computação, Rio Grande/RS, 2017.
Orientador: Dr. Leonardo Ramos Emmendorfer.

1. Agrupamento 2. Espectral 3. Particionamento 4. Aglomerativo I. Emmendorfer, Leonardo Ramos II. Título.

CDU 004

Catalogação na Fonte: Bibliotecário Me. João Paulo Borges da Silveira CRB 10/2130

Banca examinadora:

Profª. Drª. Cátia Maria dos Santos Machado

Prof. Dr. Carlos Hoppen

Prof. Dr. Leonardo Ramos Emmendorfer

AGRADECIMENTOS

Primeiramente, agradeço a Deus por estar sempre comigo me guiando nos momentos difícieis e me ajudando a superar obstáculos.

Aos meus pais, Rubens e Volma, agradeço pelo exemplo de vida que sempre foram para mim, pela luta e dedicação deles para me dar algo que ninguém pode tirar de mim, a Educação. Pelo amor e companheirismo, por terem estado presentes em todos os momentos importantes de minha vida enquanto puderam.

A toda a minha família, da qual tive apoio incondicional sempre que precisei. Pelos abraços e palavras de afeto que se tornaram importantes peças para o meu amadurecimento como pessoa.

Ao meu amor, Angélica, pela paciência, compreensão, amizade e amor que sempre me transmitiu.

Ao meu filho Nicholas, o qual ilumina cada dia de minha vida com o seu sorriso, e me impulsiona sempre a seguir em frente.

Ao meu orientador, Leonardo, por ter me guiado durante esta minha caminhada. Pela paciência, e toda a dedicação que sempre teve comigo.

As vezes é preciso aprender a correr antes de começar a andar.
— TONY STARK

RESUMO

GARCIA, Luciano Garim. **Agrupamento Espectral Aglomerativo:**
Uma Proposta de Algoritmo. 2017. 61 f. Dissertação (Mestrado) – Programa de Pós-Graduação em Computação. Universidade Federal do Rio Grande, Rio Grande.

Neste trabalho é apresentado o método de agrupamento espectral baseado em uma etapa de aglomeração dos k-menores autovetores da matriz Laplaciana, que representa o conjunto de dados a partir do grafo de similaridade. O algoritmo proposto é aplicado em diversos conjuntos de dados de formatos geométricos distintos. Os resultados são comparados aos agrupamentos obtidos pelo método k-médias e o método de agrupamento espectral via k-médias. Para medir a performance dos algoritmos é utilizada a medida-F e os resultados são apresentados em forma de tabela e gráfico. Após estudar as perfomances dos três algoritmos utilizados, conclui-se que o método apresentado neste trabalho é uma alternativa promissora ao método espectral via k-médias.

Palavras-chave: Agrupamento, Espectral, Particionamento, Aglomerativo.

ABSTRACT

GARCIA, Luciano Garim. **Agglomerative Spectral Clustering: A Proposed Algorithm**. 2017. 61 f. Dissertação (Mestrado) – Programa de Pós-Graduação em Computação. Universidade Federal do Rio Grande, Rio Grande.

In this work we present the spectral clustering based on an agglomeration step in the k-smallest eigenvectors in the Laplacian matrix, that represents the dataset from the similarity graph. The proposed algorithm is applied in many datasets of different geometric formats and the results are compared to the k-means method clustering and the k-means spectral clustering method. To measure the performance of the algorithms, the F-measure is used and the results are presented in table and graph form. After look at the performances from the three methods used, it is concluded that the presented method is an alternative approach to the k-means spectral clustering method.

Keywords: Clustering, Spectral, Partitioning, Agglomerative.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A análise de agrupamento, ou *clustering*, é o nome dado para o grupo de técnicas computacionais cujo propósito consiste em separar objetos ou indivíduos em grupos, baseando-se nas características que estes possuem. Um grupo (*cluster*) é uma coleção de objetos de dados que são similares dentro de um mesmo grupo e são dissimilares entre *clusters* distintos.

O agrupamento é a forma mais comum de aprendizado não supervisionado, ou seja, os grupos são inferidos a partir da proximidade de dados, sem indicação humana ou externa. O objetivo inicial é selecionar e extrair características dos dados e, isso muitas vezes ocorre por medidas de similaridade entre eles. A partir disso, são observados padrões de relacionamento em certos volumes, e assim, são formados agrupamentos. Os dados que possuem alto grau de dissimilaridade com os demais serão agrupados em grupos distintos. É possível verificar que a avaliação exaustiva de todas as configurações de agrupamentos possíveis é computacionalmente inviável, restringindo com isso o uso de métodos exatos para a sua solução. Dessa forma, métodos heurísticos ou aproximados têm sido propostos com frequência, os quais fornecem soluções sub-ótimas com significativa redução da complexidade na solução do problema. Entretanto, devido à grande heterogeneidade das aplicações de problemas de agrupamento, as heurísticas são normalmente desenvolvidas para determinadas classes de problemas, ou seja, não existe uma heurística que seja genérica a tal ponto que possa obter bons resultados em todas as aplicações de agrupamento.

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 muitas situações diferentes. 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 [4].

A tarefa de agrupar dados de acordo com a sua estrutura organizacional tem impulsionado diversas pesquisas na área de Aprendizagem de Máquina. Os métodos mais utilizados são os métodos de análises de componentes principais (PCA), modelos baseados em distâncias como k-médias e modelos de agrupamentos hierárquicos [12].

Neste trabalho, utiliza-se o método de agrupamento espectral baseado em particionamento de grafos para descobrir os grupos existentes em conjuntos de dados artificiais. Como alternativa de melhorar o método, é proposta uma etapa de agrupamento hierárquico aglomerativo aplicado ao mapeamento dos autovetores provenientes do polinômio característico da matriz Laplaciana do grafo de representação do conjunto.

1.1 Contextualização e Motivação

Diversos métodos de agrupamento de dados buscam encontrar agrupamentos coerentes com a sua estrutura, porém surgem limitações correspondentes a cada algoritmo utilizado. Encontrar grupos em um conjunto de dados, especialmente em dados de dimensão muito alta, é um desafio, principalmente quando os grupos possuem diferentes formatos, tamanhos e densidades, assim como dados que possuem ruído e valores de *outliers* (ver seção 5.6). Com isto, existe uma grande necessidade de elaborar métodos eficientes que realizem a tarefa de agrupar dados de maneira generalizada, independente das características do conjunto.

Um dos principais desafios é propor uma metodologia determinística ou heurística que seja capaz descobrir grupos em um conjunto de dados de modo automático. Neste sentido, métodos de agrupamento espectral vêm sendo utilizados com frequência devido à sua fácil implementação e à característica de obter um estudo global das propriedades dos dados. Esta visão global dos algoritmos baseados no método espectral é justificada pelo fato de que a similaridade entre os dados é medida par a par, desse modo, é gerada uma matriz de similaridade que contém as informações de todo conjunto.

A metodologia de agrupamento espectral proposta por [1] utiliza o algoritmo k-médias para estabelecer um agrupamento nos k-menores autovetores da matriz Laplaciana. Com isto, ao invés de agrupar um conjunto de pontos no $\mathbb{R}^n$ o método permite agrupar os dados em um conjunto no $\mathbb{R}^k$ onde k é o número de grupos correspondente aos k-menores autovetores e $k < n$. Um problema decorrente neste mapeamento é que nem sempre os autovetores possuem uma geometria simples tal que o método k-médias possa identificar facilmente os grupos corretos. Neste sentido, a utilização de um algoritmo baseado em aglomeração surge como proposta de contornar o problema em estudo.

1.2 Objetivos

Um dos propósitos deste trabalho é proporcionar ao leitor um referencial conceitual sobre o método de particionamento espectral via k-médias, além de apontar alguns problemas relacionados a esta metodologia. Problemas referentes à aleatoriedade do método e robustez quanto a parâmetros de vizinhança são os principais fatores que motivaram a escrita deste trabalho. Como contribuição de pesquisa, propõe-se adicionar ao método de agrupamento espectral um passo de agrupamento por aglomeração nos autovetores da matriz Laplaciana.

Para testar o funcionamento da metodologia com a substituição do k-médias por um algoritmo elaborado via aglomeração, objetiva-se utilizar conjuntos de dados artificiais disponíveis na literatura. A fim de medir o quão melhor uma metodologia está em relação a outra, ou vice-versa, este trabalho traz vários comparativos em forma de tabelas para esclarecer ao leitor os resultados obtidos por cada metodologia quando aplicada a algum determinado conjunto de dados.

Por fim, pretende-se que após a leitura deste trabalho, alunos, professores e pesquisadores possam utilizá-lo como referencial para futuras contribuições no estudo de métodos de particionamento espectral de grafos.

1.3 Trabalhos relacionados

Para selecionar trabalhos referentes a temática desta dissertação optou-se pela busca de artigos na literatura referente as palavras-chaves *Spectral Clustering K-means*. As bases de dados utilizadas foram IEEE Xplorer e Springer, locais onde ocorreram publicações relevantes nesta área. Em uma primeira busca ocorreram mais de 2500 resultados de artigos encontrados com as palavras-chaves informadas. Houve um processo de mineração de informação no sentido de selecionar apenas trabalhos relacionados ao tema em estudo. Foram adotados critérios de inclusão e exclusão de artigos, sendo analisados o idioma de publicação, publicações recentes, referência à Aprendizagem de Máquina, duplicidade de artigos e, principalmente, foco no objeto de pesquisa. Desse modo, o número de trabalhos estritamente relacionados foi menor do que dez, evidenciando apenas o mau funcionamento e/ou aprimoramento do algoritmo k-médias quanto utilizado pelo método espectral.

A estrutura de alguns conjuntos de dados em função de sua escala de tamanho ou densidade não permite obter um resultado satisfatório do método espectral via k-médias como cita [16]. Diversos trabalhos encontrados na literatura utilizam o k-médias como algoritmo de agrupamento no método espectral [5], [10], [3], [22], porém o método não funciona bem em conjuntos com *outliers* ou de formatos não-esféricos [9]. Para contornar o problema de escalabilidade do conjunto é proposto em [13] o cálculo de um parâmetro de escala local na medida de similaridade entre pontos vizinhos do conjunto. Além disso,

em [18] surge a proposta de uma medida de similaridade que leva em consideração a sensitividade de densidade de uma determinada região, ou seja, estabelecendo uma medida distinta da função de similaridade Gaussiana que até então era utilizada na literatura. Por fim, em [15] é apresentada uma proposta de compressão do método espectral utilizando *bandlimited graph-signals*, que tem por objetivo diminuir a complexidade do k-médias.

1.4 Organização do Trabalho

O trabalho está dividido basicamente em três partes, sendo elas, teoria, algoritmos e experimentos. O primeiro capítulo é destinado à temas introdutórios desta dissertação. No capítulo dois são apresentados ao leitor alguns conceitos teóricos de particionamento de grafos, além de uma explicação detalhada da metodologia espectral e um exemplo de aplicação. O capítulo três trata da abordagem do método de agrupamento espectral via k-médias, apresentando os algoritmos utilizados e sua complexidade computacional. De maneira similar, no capítulo quatro, é apresentado o método de agrupamento espectral aglomerativo, proposto neste trabalho. Além disso, um estudo sobre a complexidade computacional também é feito. O capítulo cinco destina-se a submeter os métodos abordados neste trabalho a diversos experimentos com conjuntos de dados artificiais, além de apresentar resultados de performance via medida-F. Por fim, no capítulo seis, é feita a conclusão deste trabalho e são apresentadas algumas sugestões de trabalhos futuros.

2 PRELIMINARES EM PARTICIONAMENTO ESPECTRAL DE GRAFOS

No que diz respeito à Teoria Espectral dos Grafos, o método de particionamento via espectro é vastamente utilizado, isto se deve ao fato de que algoritmos baseados neste contexto são de fácil implementação, além de serem razoavelmente rápidos para dados esparsos. Para compreender o funcionamento do método são apresentadas, neste capítulo, as definições e proposições referentes á Teoria de Grafos que são fundamentais para entender o funcionamento dos algoritmos espectrais utilizados neste trabalho. Ressalta-se também que não são apresentadas as demonstrações das afirmações aqui apresentadas. Porém, se o leitor julgar necessário, pode consultar todas as demonstrações disponíveis em [11].

2.1 Grafos

Para se trabalhar com agrupamento de dados via particionamento de grafos é necessário que o conceito de grafo esteja bem definido. Sendo assim, a Definição 2.1.1 tem o objetivo de esclarecer ao leitor tal conceito que será usado ao longo do texto.

Definição 2.1.1 *Um grafo $G = (V, E)$ é definido por um conjunto finito e não-vazio V, cujos os elementos são denominados vértices, e um conjunto E cujos elementos são chamados arestas. O número de vértices é indicado por $|V| = n$ e o número de arestas por $|E| = m$. Cada aresta E é definida por uma dupla $e = \{u, v\}$, tal que $u, v \in V$, e diz-se que e incide em u e v.*

O conceito de Grafo com pesos é utilizado para descrever o quão conectado um vértice está em relação ao demais, isto permite descrever sua relação com os vértices vizinhos.

Definição 2.1.2 (Grafo com Pesos) *Considere o grafo $G' = (V, E')$, tal que E' é trocado por $E = E' \times \mathbb{R}_0^+$, que considera cada aresta $e \in E'$ carregando um valor $w_{ij} > 0$. Então, $G = (V, E)$ é chamado de grafo de pesos e w_{ij} é chamado de peso de cada aresta conectando v_i e v_j.*

Para estudar certas propriedades relacionadas ao grafo, de maneira geral, é possível fazer tal tarefa por meio de matrizes que os representem. Um tipo de matriz que traz informações a respeito de um determinado grafo é a chamada matriz de adjacência, que refere-se as ligações entre os vértices do grafo.

Definição 2.1.3 (Matriz de Adjacência) *Considere um grafo $G = (V, E)$ não-orientado de pesos w_{ij}. Se $(v_i, v_j) \notin E$ define-se $w_{ij} := 0$. Então a matriz $W = (w_{ij})_{i,j=1,\dots,n}$ é chamada de matriz de adjacência de G.*

Uma maneira interessante de detectar grupos em um grafo pode ser feito a partir da análise de componentes conexos. Isto faz sentido pois se houver componentes não conectados em um grafo isto é um bom indicativo de que tais componentes não pertencem ao mesmo grupo.

Definição 2.1.4 (Componentes Conexos) *Considere um grafo $G = (V, E)$ não-orientado de pesos w_{ij} com $A \subset V$. Diz-se que é conexo se existe um caminho entre qualquer vértice de A que está completamente em A, isto é, todos os pontos do caminho estão em A, também. Além disso, diz-se que A é um componente conectado se ele é conectado e não existe nenhuma conexão com seu complementar $A^c = V - A$.*

Para analisar as relações entre cada vértice do grafo e sua ligação com os demais, a Definição 2.1.6 insere o contexto de peso de um vértice. Quanto mais peso possui um vértice dentro de um grupo, mais ele possui outros vértices relacionados consigo ou mais relação de similaridade possui com vértices próximos.

Definição 2.1.5 (Matriz de Pesos) *Define-se o peso d_i de cada vértice i como a soma dos pesos das arestas incidentes nele:*

$$d_i = \sum_{j=1}^{n} w_{ij} \tag{1}$$

A matriz de pesos do grafo G é caracterizada como uma matriz que possui todos elementos que estão fora da diagonal iguais a zero, sua forma é: $D = diag\{d_1, ..., d_n\}$.

Outro conceito a ser discutido neste trabalho é o de vetor indicador. Basicamente, sua função é indicar se dado um vértice v_i pertence ou não a um determinado grupo. Além disso, este vetor é utilizado como um dos parâmetros para estabelecer um particionamento no grafo por meio de funções de corte.

Definição 2.1.6 (Vetor Indicador) *Considere $A \subset V$ um grafo $G = (V, E)$ não-orientado de pesos w_{ij}. Define-se o vetor orientação de A por $1_A = (f_1, ..., f_n)' \in \mathbb{R}^n$ com:*

$$f_i = \begin{cases} 1 & \text{se} \quad i \in A \\ 0 & caso\, contrario. \end{cases} \tag{2}$$

Definição 2.1.7 (Grafo de Similaridade) *Considere um conjunto de dados $X = \{x_1, ..., x_n\}$ e assuma que existem valores $s_{ij} > 0$ que descrevem a similaridade de x_i e x_j para todos os pares $\{i, j\}$. Então, um grafo $G = (V, E)$ é construído considerando pontos (x_i, x_j) como vértices e as arestas representando a similaridade correspondente par a par, isto é, $w_{ij} = s_{ij}$. Assim, G é chamado de grafo de similaridade do conjunto de dados.*

Na Seção 2.2 são apresentados conceitos, definições e propriedades sobre as matrizes Laplacianas, as quais são objetos importantes de estudo neste trabalho, pois desempenham um papel fundamental no processo de agrupamento.

2.2 Matrizes Laplacianas

As matrizes Laplacianas representam os grafos de uma maneira especifíca. A Teoria Espectral de Grafos é uma área de estudo que procura extrair informações relevantes destas matrizes com o propósito de estudar algumas propriedades dos grafos. Uma informação relevante desse tipo de matriz está relacionada ao particionamento espectral de grafos, que tem por objetivo a detecção de comunidades dentro de um grafo.

Para estabelecer algumas definições importantes, considere como objeto de estudo um grafo $G = (V, E)$ não-orientado com matriz de adjacência W e matriz de pesos D.

Definição 2.2.1 *Uma matriz Laplaciana é definida como:*

$$L = D - W \tag{3}$$

A Proposição 2.2.1 enfatiza alguns propriedades da matriz Laplaciana. Tais propriedades são úteis quando se deseja estudar os autovalores e autovetores desta matriz.

Proposição 2.2.1 *Considere L a matriz Laplaciana associada com um grafo G com n vértices. Então as seguintes propriedades são válidas:*

- *Para cada vetor $(f_1, ..., f_n) = f \in \mathbb{R}^n$, tem-se*

$$f'Lf = \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2. \tag{4}$$

- *L é simétrica e positiva semi-definida.*

- *O menor autovalor de L é 0, um autovetor correspondente é o vetor constante 1.*

- *L possui n autovalores reais não-negativos $0 = \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$.*

A partir da Proposição 2.2.2 é possível relacionar o número de autovetores 0 de L com o número de componentes conexos no grafo. Este número de componentes conexos no grafo pode descrever um $k - particionamento$, a ser obtido via métodos heurísticos.

Proposição 2.2.2 *Considere G um grafo de pesos não orientado. Então a multiplicidade geométrica k de autovalores 0 de L igual ao número de componentes conectados no grafo. O auto-espaço desse autovalor é gerado pelo vetor orientação destes componentes.*

Em diversas aplicações dos métodos de particionamento espectral a utilização da matriz Laplaciana normalizada pode trazer informações mais relevantes sobre o grafo. A Definição 2.2.2 e as Proposições 2.2.3 e 2.2.4 trazem as principais propriedades relacionadas as matrizes Laplacianas normalizadas e suas implicações referentes a componentes conexos no grafo.

Definição 2.2.2 *Definem-se as matrizes Laplacianas normalizadas de acordo com as Equações (5) e (6):*

$$L_{sym} = D^{\frac{1}{2}} L D^{\frac{1}{2}} \tag{5}$$

e

$$L_{rw} = D^{-1} L. \tag{6}$$

Proposição 2.2.3 *Considere as matrizes definidas pelas Equações (5) e (6). Então as seguintes propriedades são válidas:*

- *Para cada vetor* $(f_1, ..., f_n) = f \in \mathbb{R}^n$*, tem-se:*

$$f' L_{sym} f = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} \left(\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}} \right)^2. \tag{7}$$

- λ *é um autovalor de* L_{rw} *com autovetor* u *se, e somente se,* λ *é um autovalor de* L_{rw} *com autovalor* $w = D^{\frac{1}{2}} u$*.*

- λ *é um autovalor de* L_{rw} *com autovetor* u *se, e somente se,* λ *e* u *são soluções generalizadas da Equação (8) abaixo:*

$$Lu = \lambda D u. \tag{8}$$

- 0 *é um autovalor de* L_{rw} *com o vetor constante* 1 *como o seu autovetor correspondente.* 0 *é um autovalor de* L_{sym} *com autovetor correspondente* $D^{\frac{1}{2}} 1$

- L_{sym} *e* L_{rw} *são matrizes positivas semi-definidas e possuem* n *autovalores reais não-negativos* $0 = \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$*.*

Proposição 2.2.4 *Considere* G *um grafo de pesos não orientado. Então a multiplicidade geométrica* k *de autovalores* 0 *de ambas* L_{sym} *e* L_{rw} *é igual ao número de componentes conexos* $A_1, ..., A_k$ *no grafo. Para* L_{rw}*, o auto-espaço* 0 *é gerado pelos vetores* 1_{A_i} *destes componentes. Para* L_{sym}*, o auto-espaço* 0 *é gerado pelos vetores* $D^{\frac{1}{2}} 1_{A_i}$*.*

Na seção seguinte é definido o conceito de particionamento de um grafo em k-grupos.

2.3 Particionamento de Grafos

A Definição (2.3.1) estabele o conceito de particionamento de um grafo não orientado, o qual é utilizado para a representação do conjunto de dados, conforme a Figura 1.

Definição 2.3.1 *Considere um grafo não orientado $G(V, E)$, com um conjunto de vértices $V = \{v_1, v_2, ..., v_n\}$ e um conjunto de arestas $E = \{e_1, e_2, ..., e_n\}$. O k-particionamento consiste em dividir o conjunto de vértices em k subconjuntos disjuntos $V_1, V_2, ..., V_k$ tal que $V_1 \cup V_2 \cup ... \cup V_k = V$. No caso particular em que se tem $k = 2$ acontece um bi-particionamento.*

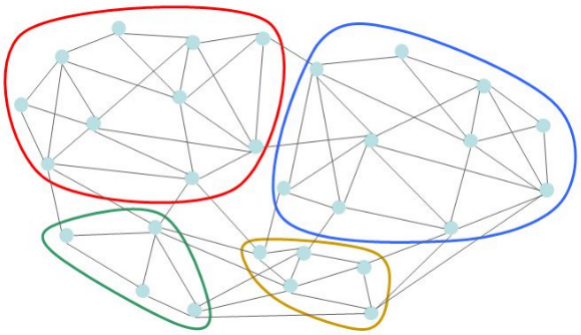


Figura 1: Grafo Particionado em 4 Grupos

Para estabelecer um particionamento no grafo que seja coerente com a sua estrutura necessita-se de métodos que façam o corte no grafo retirando os vértices que possuem menor peso, ou seja, que possuem pequena relação de similaridade. Na seção seguinte são apresentados os métodos mais comuns que executam esta tarefa.

2.4 Metodologia

Para introduzir ao leitor os passos da metodologia de agrupamento espectral, a Figura 2 apresenta um fluxograma mostrando como o método funciona pela visão do problema de particionamento de grafos. Após, é introduzido ao leitor os detalhes de cada passo que é utilizado no método, buscando fornecer aspectos práticos do funcionamento da metodologia.

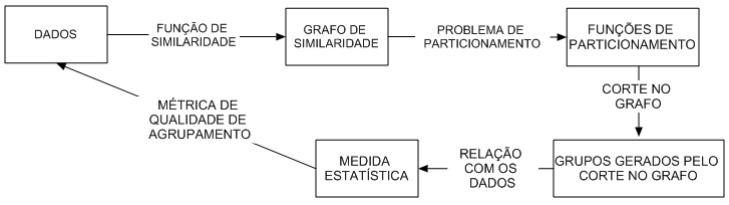


Figura 2: Fluxograma da Metodologia de Agrupamento Espectral

Considere os elementos de um conjunto de dados como os vértices de um grafo não dirigido $G = (V, E)$. O peso de cada aresta representa a similaridade entre os elementos de acordo com alguma medida [19]. O objetivo geral é particionar o grafo em dois ou

mais conjuntos disjuntos, removendo as suas arestas. Para tal tarefa, a ideia proposta é o mapeamento dos dados originais para os k menores autovetores da matriz Laplaciana obtida em função do grafo de representação dos dados, em seguida, aplica-se um algoritmo de agrupamento padrão, como o k-médias sobre estas novas coordenadas [21].

De acordo com [14] o segundo menor autovalor do Laplaciano de um grafo G, μ_{n-1}, é chamado conectividade algébrica do grafo G, sendo assim, diz-se também que um grafo é conexo se, e somente se, o seu segundo menor autovalor Laplaciano é positivo. Desse modo, o agrupamento feito pelo k-médias será elaborado de acordo com o segundo menor autovetor da matriz de autovetores, ou sobre os k menores autovetores. Assim, um conjunto de dados com dimensão n é reduzido para uma dimensão menor na qual será feito o agrupamento, reduzindo o número de iterações do k-médias se comparado à sua aplicação no conjunto original.

Para entender melhor o funcionamento da metodologia, considere um conjunto de pontos (dados) no $\mathbb{R}^n$ de acordo com a Equação (9):

$$X = \{x_1, ..., x_n | x_i \in \mathbb{R}^n\}, \tag{9}$$

onde $i = 1, ..., n$. Partindo da hipótese de que cada ponto pertencente a este conjunto possui algum grau de similaridade em relação aos demais, define-se uma medida em função de distâncias entre dois pontos e um parâmetro de similaridade, conforme a Equação (10):

$$W_{ij} = f(d(x_i, x_j), s). \tag{10}$$

Este parâmetro s pode, por exemplo, ser obtido via distribuição Gaussiana, a qual é amplamente utilizada neste contexto.

$$W_{ij} = e^{-\frac{1}{2\sigma^2} d^2(x_i, x_j)} \tag{11}$$

Apesar do extenso uso da Equação (11) é possível optar por outras medidas de similaridade, e consequentemente utilizá-las na metodologia. Algumas medidas encontradas na literatura podem ser citadas. Sejam x e y coordenadas de um ponto qualquer no $\mathbb{R}^n$, então pode-se utilizar como medidas de similaridade as Equações (12), (13), (14) e (15).

Cosseno:

$$s(x, y) = \frac{x \cdot y}{\|x\| \cdot \|y\|} = \frac{\sum_i x_i \cdot y_i}{\sqrt{\sum_i (x_i)^2} \cdot \sqrt{\sum_i (y_i)^2}} \tag{12}$$

Exponencial:

$$s(x, y) = e^{-d(x,y)}, \tag{13}$$

tal que d é a distância entre os pontos x e y.

Tanimoto:

$$s(x,y) = \frac{x^T \cdot y}{\|x\|^2 + \|y\|^2 - x^T \cdot y} \tag{14}$$

Fu:

$$s(x,y) = 1 - \frac{d_2(x,y)}{|x| + |y|}, \tag{15}$$

tal que d_2 é a distância Euclidiana.

Já analisando a construção do grafo que irá representar os dados e a relação entre eles, existem procedimentos bem variados para tal finalidade. O objetivo principal do grafo de similaridade é estabelecer uma relação de vizinhança local entre os dados. Neste trabalho são apresentados quatro metodologias.

k-nearest: Neste tipo de construção cada vértice é ligado ao seu vizinho mais próximo dependo do valor de k, que é o controle de relação local entre os vizinhos. Quando maior o valor de k, maior provavelmente será o número de vizinhos associado a um vértice. Pode-se observar um exemplo desta metodologia na Figura 3, que representa o grafo de similaridade de um conjunto de dados qualquer. Exemplos de funções que calculam distâncias entre os pontos considerando este método são:

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}, \tag{16}$$

$$\sum_{i=1}^{k}|x_i - y_i|, \tag{17}$$

$$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{\frac{1}{q}}, \tag{18}$$

onde as Equações (16), (17) e (18) representam respectivamente as distâncias Euclidiana, Manhattan e Minkowski.

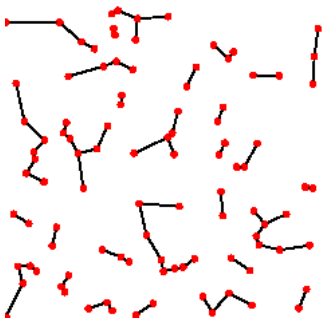


Figura 3: *k-nearest*

r-neighborhood: Nesta construção cada vértice está conectado aos vértices pertencentes a um círculo de raio r, tal que r é um valor real. Este raio deve ser definido de acordo com a vizinhança local e a quantidade de dados presentes neste local. Uma ilustração está representada na Figura 4.

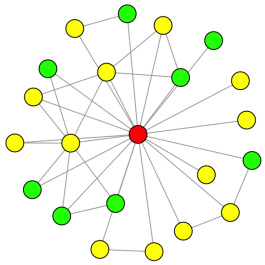


Figura 4: *r-neighborhood*

Fully-connected: Esta construção não é muito utilizada, devido ao grande número de conexões entre os vértices conforme pode ser visto na Figura 5. Neste caso, todos os vértices que possuem semelhanças não nulas são ligados uns aos outros.

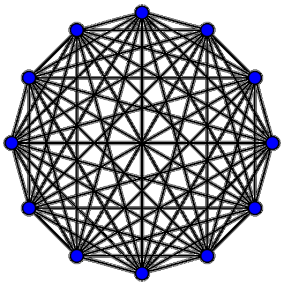


Figura 5: *Fully-connected*

k-nearest e r-neighborhood: É a combinação dos dois métodos de construção do grafo. Uma ilustração está representada na Figura 6. Tem sido amplamente utilizado.

Após obtenção do grafo que está representando os dados, é feito o seu bipariciona-mento. Nesta situação é possível determinar uma função objetivo para modelar o problema de particionamento que irá posteriormente sugerir um corte no grafo. Os próximos passos podem ser generalizados para o caso de um particionamento em k-grupos disjuntos, basta seguir analogamente a metodologia.

O objetivo geral das funções de corte é particionar um grafo $G = (V, E)$ em dois conjuntos disjuntos A e B, removendo as suas arestas. Sendo assim, a união de dois

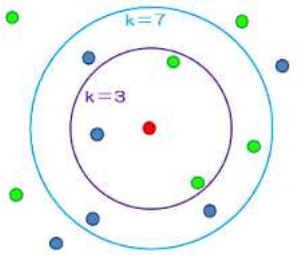


Figura 6: *k-nearest e r-neighborhood*

subconjuntos deve formar o conjunto global e a intersecção entre ambos deve ser vazia, conforme as Equações (19) e (20) respectivamente:

$$A \cup B = V \tag{19}$$

$$A \cap B = \oslash \tag{20}$$

Dados dois conjuntos disjuntos A e B de um grafo G define-se:

- O somatório dos pesos das arestas dos dois conjuntos de acordo com a Equação (21):

$$Cut(A, B) = \sum_{i \in A, j \in B} w_{ij} \tag{21}$$

- O somatório dos pesos das arestas dentro do grupo A na Equação (22):

$$Cut(A, A) = \sum_{i \in A, j \in A} w_{ij} \tag{22}$$

- O total dos pesos provenientes do grupo A conforme as Equações (1) e (23):

$$Vol(A) = \sum_{i \in A} d_i, \tag{23}$$

Com estas informações a respeito do grafo, é possível fazer o corte em G gerando dois grupos A e B. O problema está em saber se este corte realmente exclui as arestas que possuem menor peso entre os grupos, ou seja, se $Cut(A, B)$ é o menor valor considerando todas as possibilidades. Para resolver este problema utiliza-se o Método do Corte Mínimo. O principal objetivo deste método é encontrar dois grupos A e B tal que a soma dos pesos das arestas entre eles seja o menor possível. Assim, a função objetivo é dada pela Equação (24) :

$$MinCut = Cut(A, B) \tag{24}$$

Reescrevendo a Equação(24) em função da matriz de pesos e matriz de similaridade obtém-se a Equação (25):

$$MinCut = \frac{1}{4} f^T (D - W) f, \tag{25}$$

onde f é um vetor de orientação dado pela Definição 2.1.6.

A Equação (25) pode ser reescrita como:

$$(D - W)f = \lambda f, \tag{26}$$

onde D é a matriz de pesos e W a matriz de similaridade. A partir do cálculo dos autovetores dessa matriz L, o corte no grafo é feito com base no segundo autovetor de L, também conhecido como vetor de Fiedler.

Outro método de corte amplamente utilizado é o Método de Corte por Razão. Muito similar ao Método do Corte Mínimo, a função objetivo é multiplicada por um novo termo, de acordo com a Equação(27):

$$RatioCut(A, B) = Cut(A, B) \left(\frac{1}{|A|} + \frac{1}{|B|} \right). \tag{27}$$

Assim, a resolução do problema é equivalente a resolver a Equação (28) :

$$Lf = \lambda Df. \tag{28}$$

Neste caso, esta abordagem é útil visto que os agrupamentos ficam de forma balanceada. Porém, nesses métodos não são abordados as conexões dentro do grupo. Para fazer o corte no grafo é necessário que as arestas a serem cortadas tenham um valor muito pequeno, e as arestas pertencentes ao seu respectivo grupo possuam altos valores entre os vértices. Dessa maneira é feito um bom agrupamento, tratando-se em termos de similaridade, quanto mais similaridade os dados tiverem entre si melhor será o agrupamento, e quanto menos similaridade tiverem indica que os dados pertencem a grupo distintos. Para resolver este problema é necessário trabalhar-se com métodos que levem em consideração tanto as conexões fora do grupo quanto dentro do mesmo. Neste contexto, o Método do Corte Normalizado e o MinMaxCut são os mais utilizados.

Para os Métodos do Corte Normalizado e o MinMaxCut as Equações (29) e (30) modelam o problema de particionamento:

$$NCut(A, B) = Cut(A, B) \left(\frac{1}{Vol(A)} + \frac{1}{Vol(B)} \right); \tag{29}$$

$$MinMaxCut(A,B) = Cut(A,B)\left(\frac{1}{Cut(A,A)} + \frac{1}{Cut(B,B)}\right). \tag{30}$$

A minimização de $NCut$ pode ser obtida resolvendo as Equações (31) e (32). Estas equações são provenientes de uma técnica de aproximação para o mínimo de $NCut$, considerando a utilização do teorema de Rayleigh-Ritz que afirma que a solução é dada pelo segundo menor autovetor de L, também chamado de vetor de Fiedler. O teorema de Rayleigh-Ritz e as técnicas de aproximação para $NCut$ podem ser encontradas em [11].

$$(D - W)y = \lambda Dy \tag{31}$$

$$NCut = D^{-\frac{1}{2}}(D - W)D^{-\frac{1}{2}} \tag{32}$$

Desse modo, será possível obter o particionamento do grafo e separar os dados em grupos distintos. Variações destes métodos apresentados não serão abordadas neste trabalho por não serem métodos clássicos.

2.5 Aplicação da Metodologia

Para compreender melhor o funcionamento do agrupamento espectral considere o seguinte exemplo.

Dado um conjunto de dados qualquer, a primeira tarefa é montar o grafo de representação dos dados a partir da similaridade entre eles.

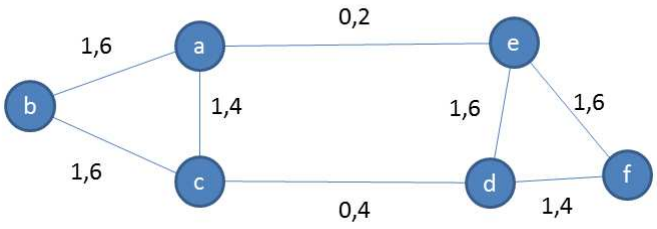


Figura 7: Grafo de Representação dos Dados

A partir da análise da Figura 7 é possível inferir que existem dois possíveis grupos, tal que seus vértices possuem um alto grau de similaridade entre si. Porém, em conjunto de dados muito grandes estes possíveis agrupamentos não ficam tão claros comparados a este exemplo. Assim, é necessário utilizar o algoritmo de agrupamento espectral, extraindo a matriz de similaridade a partir do grafo criado.

Matriz de Similaridade						
Vértices	a	b	c	d	e	f
a	0	1,6	1,2	0	0,2	0
b	1,6	0	1,6	0	0	0
c	1,2	1,6	0	0,4	0	0
d	0	0	0,4	0	1,6	1,4
e	0,2	0	0	1,6	0	1,6
f	0	0	0	1,4	1,6	0

Após a obtenção da matriz de similaridade, é calculada a matriz de pesos com a finalidade de definir a matriz Laplaciana a ser utilizada para efetuar o agrupamento. A montagem da matriz de pesos é proveniente das Equações (23) e (2.1.5), fazendo com que os elementos da diagonal principal sejam os únicos a serem não nulos.

Matriz de Pesos					
3	0	0	0	0	0
0	3,2	0	0	0	0
0	0	3,2	0	0	0
0	0	0	3,4	0	0
0	0	0	0	3,4	0
0	0	0	0	0	3

Obtidas as matrizes W e D, é feita a diferença $L = D - W$ e, com isso, são calculados os autovalores desta matriz L.

Matriz Laplaciana					
3	-1,6	-1,2	0	-0,2	0
-1,6	3,2	-1,6	0	0	0
-1,2	-1,6	3,2	-0,4	0	0
0	0	-0,4	3,4	-1,6	-1,4
-0,2	0	0	-1,6	3,4	-1,6
0	0	0	-1,4	-1,6	3

Calculando as aproximações para os autovalores desta matriz, obtém-se:

λ
0
0,6
4,4
4,6
5
6

Após obtêm-se os respectivos autovetores de L:

Matriz de Autovetores					
0,8	0,4	0,2	0,8	-0,4	-1,8
0,8	0,4	0,2	0	0,8	0,6
0,8	0,4	-0,4	0	-0,4	1,2
0,8	-0,8	1,8	0,4	-0,8	-1,2
0,8	-1,4	-0,8	-1,6	-1,2	-0,4
0,8	-1,4	-0,4	1	1,6	1,8

O critéro para escolher o autovetor em que será feito o agrupamento é baseado na Teoria de Fiedler, que sugere que a partição seja estabelecida considerando o segundo menor autovetor. Desse modo, o autovetor é extraído da matriz e usado para gerar os grupos:

Vetor de Fidler	
a	0,4
b	0,4
c	0,4
d	-0,8
e	-1,4
f	-1,4

O que este autovetor sugere é uma partição do grafo em dois grupos:

*__Grupo A__- Formado pelos vértices a, b e c, provenientes do mapeamento do segundo autovetor de valores $0,4$.

*__Grupo B__- Formado pelos vértices d, e e f, provenientes do mapemento do segundo autovetor de valores $-0,8$, $-1,4$ e $-1,4$.

Desse modo, o novo grafo é representado pela Figura 8 :

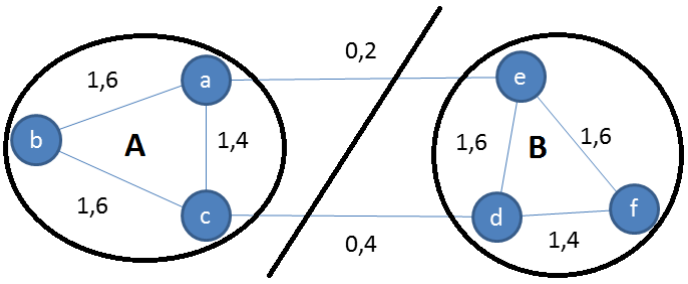


Figura 8: Grafo Particionado

Uma vantagem desta metodologia é a aplicação de um algoritmo de agrupamento diretamente nos autovetores provenientes da matriz Laplaciana. A vantagem computacional está relacionada a diminuição da dimensão de um conjunto de dados em 1 ou k coordenadas, dependendo da abordagem utilizada. Na prática, um algoritmo como o k-médias, amplamente usado, pode diminuir o número de iterações necessárias para a sua convergência quando aplicado apenas ao conjunto de autovetores. Esta eficiência tende a aumentar dependendo diretamente da dimensão dos autovetores, ou seja, no caso em que o agrupamento ocorre apenas no segundo autovetor existe somente uma coordenada para o k-médias atualizar os centros dos grupos.

3 ABORDAGEM DE AGRUPAMENTO ESPECTRAL VIA K-MÉDIAS

Neste capítulo são apresentados os algoritmos k-médias e espectral via k-médias e suas respectivas características. A seção final trata da análise de complexidade computacional de tempo da metodologia espectral.

3.1 Algoritmo K-médias

O algoritmo k-médias baseia-se na minimização de uma medida de custo, a distância interna entre os padrões de um agrupamento. 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-médias

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

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

1. Escolha estimativas iniciais arbitrárias $\theta_j(0)$ para os $\theta_{j's}, j = 1, ..., m$ (i.e. para os centróides dos k clusters);
2. Repita:

for $i \leftarrow 1$ **to** N **do**
    Determine o representante mais próximo, isto é, θ_j (centróide 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 Algoritmo 1 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-médias pode não funcionar de maneira esperada dependendo da curvatura da distribuição dos dados no plano bidimensional. Geralmente, o método funciona bem em conjuntos onde os grupos estão bem separados, ou seja, em distribuições gaussianas.

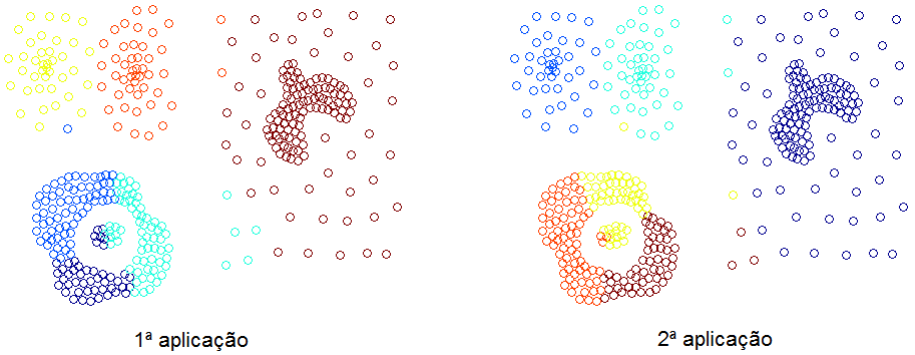


Figura 9: Aplicação do Algoritmo K-médias Direto no Conjunto de Dados

Na Figura 9 está um exemplo de aplicação do algoritmo k-médias no conjunto de dados *Compound* proveniente do repositório de Aprendizagem de Máquina UCI. Este exemplo ressalta a dificuldade do método em trabalhar com conjunto de dados de formato não esférico. Além disso, nota-se que os resultados de agrupamento do mesmo conjunto de dados com o k-médias não foram os mesmos, evidenciando a necessidade de uma boa inicialização do método.

Na seção seguinte é apresentado o algoritmo de agrupamento espectral via k-médias. Visto que apenas a utilização do Algoritmo 1 não é o ideal em vários casos, então o algoritmo espectral tem por objetivo apresentar um novo conjunto de dados que represente o anterior e seja mais simples para a aplicação do k-médias.

3.2 Algoritmo Espectral via K-médias

O algoritmo pioneiro desta metodologia surgiu graças ao trabalho de [1] que apresentou sua metodologia no artigo intitulado "*On spectral clustering: Analysis and an Algorithm*". Os passos do Algoritmo 2 basicamente desempenham a tarefa de obter a matriz Laplaciana do grafo de similaridade e calcular os seus autovetores para agrupá-los via k-médias.

O parâmetro de escala σ^2 (desvio padrão) no Algoritmo 2 controla a rapidez com que a matriz de similaridade W decresce levando em consideração a distância entre x_i e x_j [1]. A matriz D 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

Algoritmo 2: Agrupamento Espectral via K-médias [1]

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

Saída : Grupos $A_1, ..., A_k$

1.Formar a matriz de similaridade W definida por: $W_{ij} = e^{\left(-\frac{1}{2\sigma^2}d^2(x_i,x_j)\right)}$.
2.Construir a matriz Laplaciana normalizada simétrica: $L = D^{-\frac{1}{2}}(D - W)D^{-\frac{1}{2}}$.
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: $U \left[u_1 \vdots \cdots \vdots u_k\right] \in \mathbb{R}^{n\times k}$
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 k-médias. $Y_{ij} = \frac{U_{ij}}{\left[\sum_{j=1}^{n} U_{ij}^2\right]^{\frac{1}{2}}}$.
6.Colocar os pontos originais x_i no grupo j, se e somente se, a linha i da matriz Y for colocada no cluster j.

matriz Laplaciana.

Calculando-se os k autovetores de L, obtém-se as matrizes U e logo após Y, onde no passo 5 ocorre a aplicação do Algoritmo 1. Ressalta-se que, nesta dissertação, o passo 5 não seja mais utilizado, ou seja, propõe-se a sua substituição por outra metodologia que possa ser alternativa a esta utilizada pelo Algoritmo 2.

Na seção seguinte é feito um estudo sobre a complexidade computacional, principalmente, do Algoritmo 2, analisando sua complexidade de tempo.

3.3 Complexidade Computacional

Para medir o custo de execução de um algoritmo é comum definir uma função de custo ou função de complexidade f. A função de complexidade de tempo: $f(n)$ mede o tempo necessário para executar um algoritmo para um problema de tamanho n. Na realidade, a complexidade de tempo não representa tempo diretamente, mas o número de vezes que determinada operação considerada relevante é executada. Para testar os limites do algoritmo quanto à sua complexidade computacional é interessante estudar o seu comportamento no pior caso, ou seja, $f(n) = O(f(n))$. Quando a notação O é usada para expressar o tempo de execução de um algoritmo no pior caso, está se definindo também o limite (superior) do tempo de execução desse algoritmo para todas as entradas.

Para estudar a complexidade de tempo do Algoritmo 2 é necessário definir o número de operações que são executadas em cada etapa do método. Para estudar o pior caso considera-se que o algoritmo execute todas as operações possíveis levando em consideração o número total de entradas. Além disso, vale ressaltar que a complexidade

de tempo deste método está diretamente ligada ao número de elementos e a dimensão do conjunto de dados utilizado como entrada no algoritmo.

A construção da matriz de similaridade do grafo de representação dos dados é considerada uma das tarefas com maior custo dentro do algoritmo. Para elemento de W_{ij} é calculado no mínimo um produto de x_i por x_j, podendo esta operação se estender devido a dimensão d do conjunto. Ou seja, para esta etapa o custo de tempo é:

$$O\left(n^2 d\right). \tag{33}$$

Para fazer W_{ij} uma matriz esparsa, emprega-se a abordagem dos k-vizinhos mais próximos e retem-se somente W_{ij} onde i (ou j) está entre os k-vizinhos mais próximos de j (ou i). Escaneando uma vez W_{ij} para $j = 1, ..., n$ e mantendo um montante máximo com tamanho k, inserimos sequencialmente a similaridade que é menor do que o valor máximo do montante. Assim, a complexidade para um ponto x_i é $O(n \log k)$ uma vez que a reestruturação de um montante máximo está na ordem de $\log k$. Segundo [24] a complexidade para tornar W uma matriz esparsa é:

$$O(n^2 \log k). \tag{34}$$

Para calcular os autovetores da matriz Laplaciana utiliza-se, geralmente, um solver já implementado. Para tal procedimento a complexidade está também relacionada a quantidade de elementos do conjunto de dados. Neste caso, a complexidade de tempo é dada por:

$$O(n^3). \tag{35}$$

A aplicação do k-médias nos resultados de custos de decomposição de autovalores é:

$$O(nldk), \tag{36}$$

onde n é o número de pontos de dados de entrada, l é o número de iterações de k-médias, d é a dimensionalidade dos dados de entrada e k é o número de grupos finais. Segundo [23], apesar da importância dos algoritmos de agrupamento espectral, eles não são amplamente vistos como um concorrente para algoritmos clássicos como agrupamento hierárquico e k-médias para problemas de mineração de dados em larga escala. A complexidade computacional geral do algoritmo de agrupamento espectral é $O(n^3)$. Isso torna os métodos de agrupamento espectral inviáveis para problemas com n na ordem de milhares. Além disso, problemas com n na ordem de milhões (ou bilhões) estão inteiramente fora de alcance.

4 ABORDAGEM DE AGRUPAMENTO ESPECTRAL AGLOMERATIVO

A proposta de elaboração de um algoritmo de agrupamento espectral sem a etapa de mapeamento dos autovetores utilizando o k-médias é motivada pelos resultados aleatórios que o método produz. Apesar da convergência do k-médias ocorrer em poucas iterações, uma má inicialização dos centroides é um problema que afeta o resultado do agrupamento. Isto ocorre porque os centroides são inicializados em posições aleatórias no espaço dimensional dos dados, influenciando no resultado final do agrupamento. Se existem k grupos reais, então a probabilidade de selecionar um centroide para cada grupo é relativamente pequena [20].

Para o método k-médias é possível utilizar uma equação que calcula erros em função das distâncias entre os elementos e seus grupos, dada pela Equação (4):

$$E(x) = d(x, c_i), \tag{37}$$

onde $E(x)$ representa a distância de x até o centróide c_i de seu grupo C_i.

O principal objetivo do k-médias é minimizar a soma dos erros (SSE = sum of square errors), dada pela Equação

$$SSE = \sum_{i}^{k} \sum_{x \in C_i} d(x, C_i)^2. \tag{38}$$

Nem sempre o k-médias consegue encontrar o mínimo para o SSE, visto que, a minimização deste erro depende muito da escolha dos centroides iniciais. Para uma escolha aleatória têm-se diversas execuções do k-médias até encontrar a escolha que produz o menor SSE. Vale lembrar também que nem sempre a execução do algoritmo de agrupamento é suficiente para obter bons resultados, pois a qualidade do agrupamento depende da distribuição dos dados e do número k de grupos que se quer encontrar. Sendo assim, o método espectral pode gerar autovetores que correspondem corretamente com a estrutura organizacional dos dados, porém se o algoritmo que for utilizado para agrupar os autovetores não for eficaz, então o método falha.

Além disso, a metodologia espectral via k-médias é sensível ao parâmetro de vizinhança que é estabelecido na construção do grafo de similaridade. Na Figura 10 é possível identificar três configurações distintas do grafo de similaridade, onde cada representação indica a criação do grafo de acordo uma vizinhança fixa. Desse modo, o método espectral pode ”enxergar”o mesmo conjunto de dados com uma ”visão”mais local ou mais global, influenciando diretamente no resultado dos autovetores, onde é feito o agrupamento.

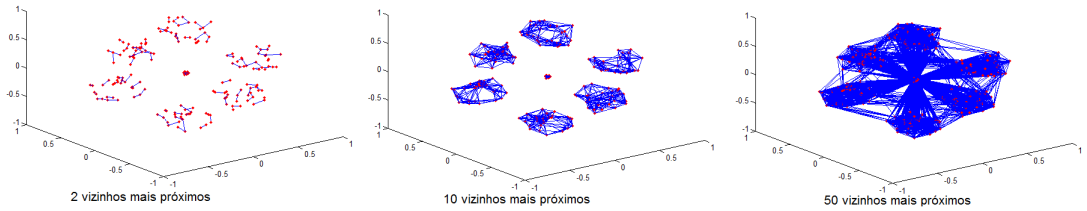


Figura 10: Representação do Grafo de Similaridade para um Conjunto de Dados Considerando Variações no Número de Vizinhos

Na seção 5.3 do capítulo 5 é descrito tal característica associada a um exemplo, onde o número de vizinhos mais próximos informado aos algoritmos não é o mais apropriado para a criação do grafo de similaridade. Desse modo, a distribuição dos elementos dos autovetores no plano não possui uma geometria simples para a aplicação do k-médias. Sendo assim, neste trabalho buscou-se uma abordagem por aglomeração na etapa de mapeamento dos autovetores para que possa se tornar um método alternativo para contornar os problemas existentes com a abordagem via k-médias. Uma das vantagens do método proposto é que o resultado de agrupamento não possui variação, ou seja, não é necessário executar o algoritmo diversas vezes para se obter um bom resultado. Além disso, o método consegue convergir, em muitos casos, com apenas uma ou duas iterações em casos de biparticionamento. Na próxima seção, são descritos alguns conceitos de algoritmos aglomerativos e como funcionam. Após, o conceito de aglomeração é incorporado ao método espectral na etapa de mapeamentos dos autovetores.

4.1 Algoritmo Aglomerativo

Na classe de algoritmos hierárquicos existem duas categorias:

Algoritmos Divisivos- Nesta categoria a abordagem de agrupamento é chamada de *top-down*, ou seja, de cima para baixo. Dado um conjunto de dados pode-se estabelecer um agrupamento com relação ao conjunto global levando em consideração toda a informação de similaridade proveniente do conjunto. De maneira recursiva pode-se estabelecer quantos grupos se queira até a convergência do método. Em conjuntos de dados de formato hiperesférico e bem separados esta abordagem é interessante, como pode-se

ver na Figura 11:

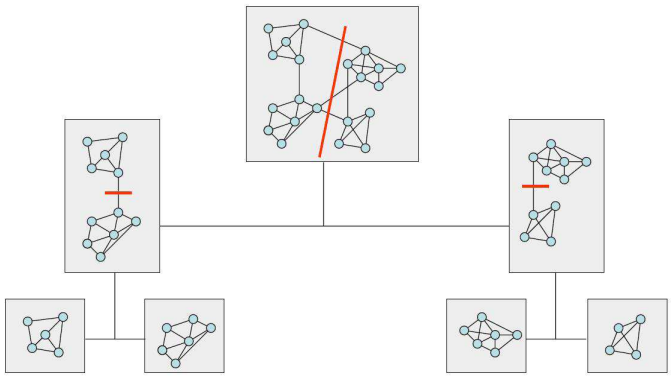


Figura 11: Abordagem *Top-Down*

Algoritmos Aglomerativos- Quando é necessário agrupar dados que não possuam uma geometria simples, então os métodos divisivos não são uma boa escolha. Para obter informações locais na vizinhança de cada elemento de um conjunto a abordagem *botton-up* é uma alternativa. Nesta categoria, os métodos aglomerativos destacam-se por estabelecerem agrupamentos locais até obter apenas um ou mais conjuntos globais. Assim, optou-se neste trabalho por utilizar o método de agrupamento aglomerativo, pois este é capaz de obter informações organizacionais do conjunto de maneira local, obedecendo a geometria dos dados.

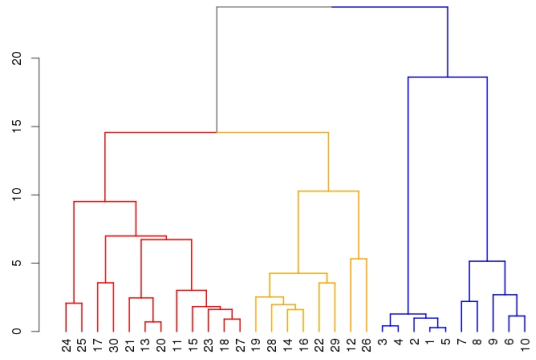


Figura 12: Árvore Hierarquica Aglomerativa

Basicamente, o método pode ser representado por meio de uma árvore de hierarquias conforme Figura 12. Inicialmente o algoritmo considera cada elemento do conjunto como um grupo isolado. Por iteração cada grupo vai se unindo a outros de acordo com a sua distância em relação a eles. Sendo assim, é indispensável a utilização de uma medida

de distância que calcule a dissimilaridade entre cada grupo. Neste caso, existem quatro métodos muito utilizados para calcular distância entre os grupos, conforme Tabela 1:

Tabela 1: Distância entre Grupos Usados em Diferentes Algoritmos Aglomerativos

Método	Distância entre Grupos
Single-link	$d(C_a, C_b) = min_{ij} d(x_i \in C_a, x_j \in C_b)$
Complete-link	$d(C_a, C_b) = max_{ij} d(x_i \in C_a, x_j \in C_b)$
Average-link	$d(C_a, C_b) = \frac{1}{\|C_a\|\|C_b\|} \sum_{x_i \in C_a} \sum_{x_j \in C_b} d(x_i, x_j)$
Centroid-link	$d(C_a, C_b) = d(\overline{C_a}, \overline{C_b})$

No método *Single-link* os elementos mais próximos são agrupados juntos, ou seja, neste caso o algoritmo não utiliza nenhum critério global para estabelecer os grupos. Neste tipo de abordagem pode ocorrer a formação de grupos muito grandes, em outras palavras significa dizer que a distância dentro de conjuntos com baixa densidade pode ser maior do a distância entre dois grupos reais, causando um agrupamento equivocado. Uma maneira de contornar tal problema pode ser a partir da utilização do método *Complete-link*, que basea-se na informação global do conjunto, assim, não havendo a preocupação com a formação de grupos grandes conforme o método de conexão única. Apesar de tal funcionalidade este método pode apresentar sensibilidade a *outliers*, desconfigurando a formação correta dos grupos. No método *Average-link* que calcula a média entre os grupos, os problemas com *outliers* e grupos grandes são desconsiderados, pois este critério estuda tanto o comportamento local quanto global do conjunto de dados. Por fim, o método *Centroid-link* é promissor quando os grupos são globulares e de mesma densidade, ou seja, tal critério funciona bem apenas em um conjunto limitado de problemas.

O Algoritmo 3 abaixo foi escrito de maneira autoral, podendo ser encontrado na literatura com variações em seus passos. Se a matriz Y for a matriz de distâncias do conjunto de dados original, então o método pode ser aplicado puramente, sem a necessidade do uso dos autovetores. Porém, neste trabalho, os autovetores são utilizados como informação prévia para a aplicação do método aglomerativo, desse modo, o método espectral é indispensável para obter bons resultados com o Algoritmo 3.

A busca pelo número de grupos que acontece nos autovetores é feita de maneira aglomerativa dentro desse novo conjunto . A informação que os autovetores carregam do conjunto de dados é utilizada como um rótulo de agrupamento, ou seja, cada elemento de um autovetor possui um número associado a ele, que é utilizado no processo de aglomeração.

Uma matriz de distâncias é gerada de acordo com as coordenadas dos autovetores, e logo após é determinada a média das distâncias. São inicializados três vetores, o primeiro contendo o número inicial de rótulos de acordo com o tamanho dos autovetores, o vetor distância obtido de acordo com a média das distâncias e o vetor número de vizinhos inicializado com todas posições iguais a um. São selecionadas duas posições aleatórias

Algoritmo 3: Algoritmo Aglomerativo

Entrada: k-menores autovetores normalizados: Matriz Y
Saída : Grupos rotulados $A_1, ..., A_k$

1.Construir a matriz de distância Euclidiana dos elementos de Y, $dist(y_i, y_j) = \sqrt{\sum_{i=1}(y_i - y_j)^2}$;
2.Calcular a média das distâncias da matriz Y nas posições a e b, ou seja, $d(Y_a, Y_b) = \frac{1}{|Y_a||Y_b|} \sum_{y_i \in Y_a} \sum_{y_j \in Y_b} dist(y_i, y_j)$;
3.Inicializar os vetores rótulos, números de vizinhos e distância entre os vizinhos;
$rot \leftarrow (1, 2, ..., n)^t$
$numviz \leftarrow (1, ..., 1)^t$
$dviz \leftarrow d(Y_a, Y_b).(1, ..., 1)^t$
4.Tomar duas posições quaisquer a e b dos vetores inicializados em 3 e atualizar seus itens de acordo com a distância média entre os grupos(rótulos);
if $rot(a) \neq 0$ *and* $dist(a, b) \cdot numviz(a) <= dviz(a)$ **then**
    $dviz(b) += dist(a, b)$;
    $numviz(b) += 1$;
    $rot(b) = rot(a)$;
if $numviz(b) >= viz$ **then**
    $rot(b) = 0$;
    $numviz(b) = 1$;
    $dviz(b) = d(Y_a, Y_b)$;
5.Parar o processo de aglomeração quando todas as posições dos vetores de 3 forem submetidos aos testes em 4.

do conjunto dos autovetores, onde cada posição passará por um teste de distância no grupo de acordo com o seu rótulo, o número de vizinhos próximos e distância entre a outra posição aleatória.

Se o rótulo da posição a for diferente de zero e o número de vizinhos multiplicados pela distância entre as posições a e b forem menores que a média da distância da posição a, então o rótulo da posição b torna-se o mesmo da posição a, o número de vizinhos de b incrementa em uma unidade assim como a média da distância da posição b incrementa de acordo com a distância entre a e b.

Como opção parar o processo de aglomeração em um posição é utilizada uma comparação entre o número de vizinhos desta posição e o número de vizinhos do conjunto de dados atual (autovetores) representado pela variável viz. Se a quantia de vizinhos da posição extrapolar o número de vizinhos do conjunto de dados, então o processo de aglomeração naquela posição termina. A parada total do método ocorre quando os testes do passo 4 são executados para todas as posições dos vetores em 3.

Após o processo de aglomeração cada elemento de um determinado grupo possui um rótulo, que é a sua característica em relação a todo o conjunto. Na Figura 13 é apresentado um exemplo no grafo da distribuição dos rótulos em três grupos distintos. Espera-se

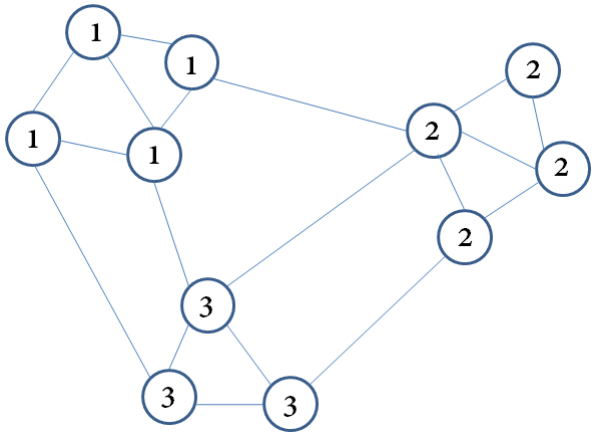


Figura 13: Grupos rotulados

que ao final da execução do Algoritmo 3, a quantidade de rótulos com um determinado valor seja igual ao número de grupos do conjunto. Assim, cada rótulo é associado as coordenadas correspondentes de cada elemento formando os grupos.

Na próxima seção, o Algoritmo 3 é acoplado ao método espectral com a intenção de substituir o passo 5 do Algoritmo 2.

4.2 Algoritmo Espectral Aglomerativo

Na construção do Algoritmo 4 foi utilizado uma parte do Algoritmo 2 proposto por [1]. Agora, nesta versão, os autovetores não são mais agrupados de acordo com metodologia baseada em centroides. Propõe-se, o conceito de aglomeração de grupos via distância média entre si. Vale ressaltar que a distância entre grupos também pode ser medida de acordo com outras métricas, como por exemplo, as métricas apresentadas na Tabela 1. Como a forma de medir distâncias entre grupos pode variar os resultados de agrupamento, optou-se, neste trabalho, apenas por utilizar a distância média, desse modo, todos os experimentos e testes mostrados aqui foram feitos com esta medida.

Os passos 6 até 10 no Algoritmo 4 possuem a tarefa de formar os grupos de acordo com as informações de rótulo, vizinhança e distância dos pontos (grupos). Este processo pode-se repetir iterativamente até, por exemplo, existir apenas um único grupo. Isto ocorre devido a natureza do método hierárquico aglomerativo, porém, para se obter um agrupamento coerente com a sua estrutura apenas uma iteração do método aglomerativo é o suficiente. Em conjunto de dados que possuem vários grupos, mais que dez, é interessante fazer duas ou três iterações do método.

Na seção seguinte é apresentado a complexidade computacional de tempo referente ao Algoritmo 4.

Algoritmo 4: Algoritmo Espectral Aglomerativo

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

Saída : Grupos rotulados $A_1, ..., A_k$

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

2.Construir a matriz Laplaciana normalizada simétrica:
$L = D^{-\frac{1}{2}}(D - W)D^{-\frac{1}{2}}$;

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: $U \left[u_1 \vdots \cdots \vdots u_k\right] \in \mathbb{R}^{n\times k}$;

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 $\mathbb{R}^k$, isto é, $Y_{ij} = \frac{U_{ij}}{\left[\sum_{j=1}^{n} U_{ij}^2\right]^{\frac{1}{2}}}$.

6.Construir a matriz de distância Euclidiana dos elementos de Y,
$dist(y_i, y_j) = \sqrt{\sum_{i=1}(y_i - y_j)^2}$;

7.Calcular a média das distâncias da matriz Y para as posições a e b, ou seja,
$d(Y_a, Y_b) = \frac{1}{|Y_a||Y_b|}\sum_{y_i\in Y_a}\sum_{y_j\in Y_b} dist(y_i, y_j)$;

8.Inicializar os vetores rótulos, números de vizinhos e distância entre os vizinhos;

$rot \leftarrow (1, 2, ..., n)^t$

$numviz \leftarrow (1, ..., 1)^t$

$dviz \leftarrow d(Y_a, Y_b).(1, ..., 1)^t$

9.Tomar duas posições quaisquer a e b dos vetores inicializados em 8 e atualizar seus itens de acordo com a distância média entre os grupos(rótulos);

if $rot(a) \neq 0$ and $dist(a, b) \cdot numviz(a) <= dviz(a)$ **then**
- $dviz(b) += dist(a, b)$;
- $numviz(b) += 1$;
- $rot(b) = rot(a)$;

if $numviz(b) >= viz$ **then**
- $rot(b) = 0$;
- $numviz(b) = 1$;
- $dviz(b) = d(Y_a, Y_b)$;

10.Parar o processo de aglomeração quando todas as posições dos vetores de 8 forem submetidos aos testes em 9.

4.3 Complexidade de Computacional

Para estudar a complexidade de tempo do algoritmo espectral aglomerativo é necessário apenas considerar as medidas de complexidade obtidas pelo Algoritmo 2 até a etapa de obtenção dos autovetores da matriz Laplaciana e soma-las às medidas obtidas pela etapa de aglomeração.

Sabe-se da análise feita anteriormente que o processo que exige mais tempo computacional é dado pelo cálculo dos autovetores, gerando uma complexidade de tempo na ordem de $O(n^3)$. Após esta etapa ocorre a aplicação do algoritmo de aglomeração aplicado ao conjunto de autovetores de tamanho n x k, onde n é a quantidade de elementos do conjunto original de dados e k é o número de menores autovetores da matriz Laplaciana. Quando o agrupamento ocorre apenas no segundo menor autovetor (Vetor de Fiedler) o conjunto de autovetores irá medir n x 1.

A complexidade de tempo do agrupamento aglomerativo é:

$$O(n^2 \log n). \tag{39}$$

Primeiro calcula-se todas as distâncias n^2 para os grupos iniciais, e aglomera os elementos em grupos (tempo: $O(n^2 \log n)$). Em cada uma das $O(n)$ iterações, identifica-se o par de grupos com a coesão mais alta em $O(n)$; Agrupa-se o par; e atualiza-se as médias dos grupos. Para cada grupo, também atualiza-se a lista ordenada de elementos, excluindo os dois grupos já aglomerados. Assim, cada iteração toma $O(n \log n)$. A complexidade do tempo neste caso é $O(n^2 \log n)$.

Estudando o número de operações executadas no passo de aglomeração conclui-se que no geral a complexidade do método ainda continua sendo $O(n^3)$, devido ao esforço computacional exigido no cálculo dos autovetores de matriz Laplaciana. Em geral, para aplicação do método em grandes bancos de dados o algoritmo necessitaria de uma otimização nas etapas que possuem alto custo de execução.

5 EXPERIMENTOS E TESTES

Neste capítulo são apresentados uma série de experimentos feitos com a aplicação de três algoritmos de agrupamentos em diversos conjuntos de dados. Além dos algoritmos espectrais abordados neste texto, a utilização do k-médias tem por objetivo obter uma comparação de resultados em relação aos outros dois. A escolha da utilização do k-médias nos experimentos foi feita de acordo com a sua popularidade em diversas aplicações, sendo relevante sua comparação de resultados em relação aos métodos espectrais. 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.

5.1 Medida de Qualidade de Agrupamento

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

$$F(y, h(x)) = \frac{(1+\beta^2)\sum_{i=1}^{m} y_i h_i(x)}{\beta^2 \sum_{i=1}^{m} y_i + \sum_{i=1}^{m} h_i(x)} \in [0, 1], \tag{40}$$

onde $\frac{0}{0} = 1$. Esta medida essencialmente corresponde a uma média harmônica ponderada de precisão e exaustividade [8]. 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.

O principal objetivo de utilizar a medida-F é relacionar os resultados obtidos pelo agrupamento espectral aglomerativo com um parâmetro estatístico. Os resultados de performance obtidos pelos outros dois algoritmos servem apenas para comparação, enfatizando em qual situação são mais eficazes ou não.

5.2 Experimentos 1

Para a aplicação dos métodos de agrupamento discutidos aqui, foram escolhidos alguns conjuntos de dados disponíveis em [6], [7] e [2]. 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. Em cada teste são previamente informados pelo usuário a quantidade de grupos a serem formados, uma tabela contendo as coordenadas x e y de cada elemento do conjunto, o desvio padrão geral dos dados e o número desejado de vizinhos mais próximos para a construção do grafo de similaridade.

No primeiro teste foi utilizado o conjunto de dados *Jain* que possui o formato de duas meia luas conforme a Figura 14. 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 relacionado 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 irão 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 14.

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 14. Para montar o grafo de similaridade deste conjunto é interessante que se estabeleca uma vizinhança pequena, na qual possa representar de maneira adequada a vizinhança local, facilitando a aplicação dos algoritmos.

Conforme a visualização da Figura 14 é possível notar que o algoritmo k-médias 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-médias neste caso não é o mais recomendado. Já os grupos formados pelo algoritmo espectral via k-médias gerou 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. Os parâmetros informados aos métodos foram os seguintes: desvio padrão com $\sigma = 10.3649$, número de grupos $k = 2$ e número de vizinhos $num\_neighbour = 10$. Pode-se notar que os resultados do último algoritmo de agrupamento foram satisfatórios, respeitando a estrutura organizacional dos dados. A explicação para este fato está relacionada ao método utilizado, visto que, o mesmo busca obter localmente uma relação entre os elementos de acordo com a sua

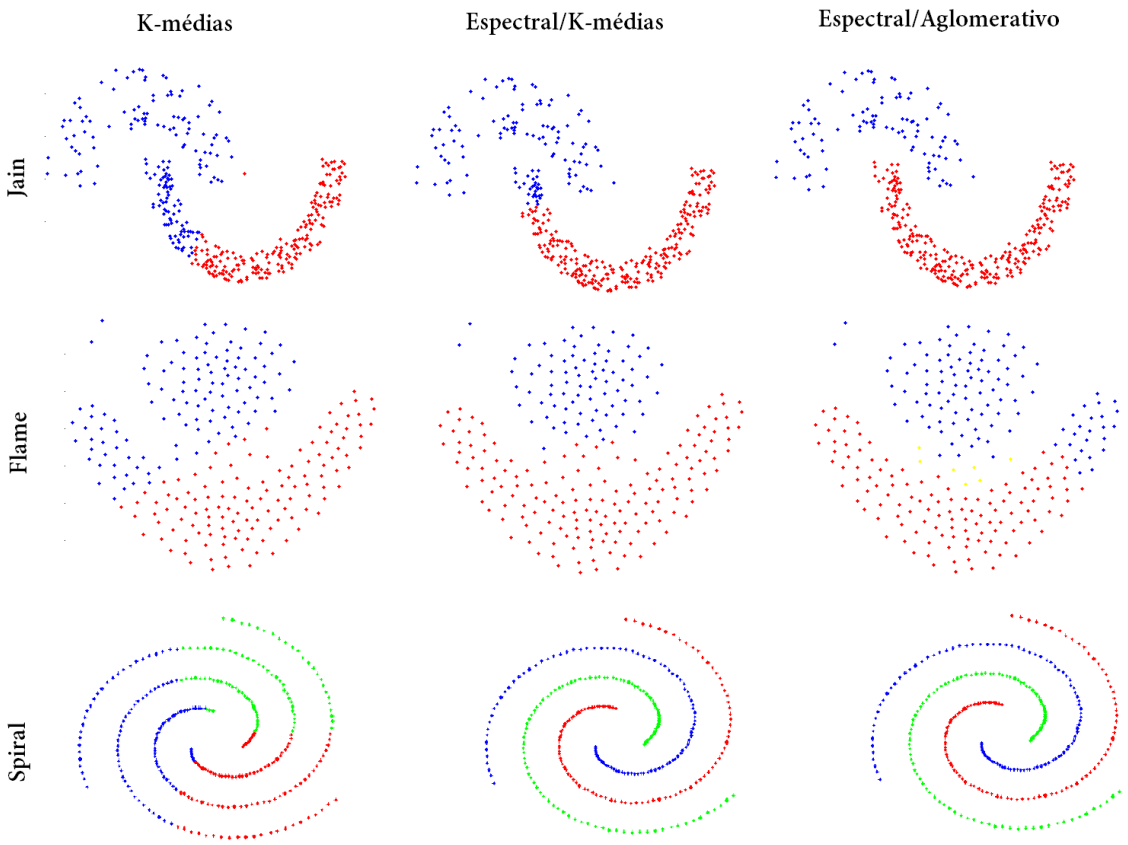


Figura 14: Resultados dos Experimentos Referentes aos Conjuntos de Dados *Jain*, *Flame* e *Spiral* para os Algoritmos K-médias, Espectral/K-médias e Espectral/Aglomerativo.

média dentro do conjunto de autovetores.

No agrupamento do conjunto *Flame* estabelecido pela metodologia k-médias, os grupos se formaram 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-médias consegue eficientemente colocar um centroide correto nesta região. Porém, em relação à meia lua este acerto do método torna-se impossível. Os parâmetros utilizados foram: desvio padrão com $\sigma = 7.5630$, número de grupos $k = 2$ e número de vizinhos $num\_neighbour = 5$. 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. No processo de aglomeração dos autovetores o resultado obtido foi bom, enfatizando em boa parte a distinção entre o grupo de formato esférico e a meia lua.

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-médias o agrupamento do conjunto *Spiral* não foi satisfatório, novamente o problema ocorre devido a curvatura da distribuição dos elementos no plano. Diferente de muitos casos, neste conjunto nem uma boa inicialização do centroide irá estabelecer corretamente os grupos, sendo assim, o algoritmo é ineficaz neste caso. Em contrapartida, os algoritmos baseados na metodologia de particionamento espectral apresentaram bons resultados. As entradas dos algorimtos foram de acordo com as seguintes informações: desvio padrão com $\sigma = 7.1561$, número de grupos $k = 3$ e número de vizinhos $num\_neighbour = 2$. Pode-se notar na Figura 14 o agrupamento nos autovetores ocorreu de maneira esperada, onde o mapeamento no conjunto de dados foi o melhor possível. Neste teste não houve diferença entre o método espectral via k-médias e o método espectral aglomerativo, pois a estrutura organizacional dos autovetores estava bem definida.

Tabela 2: Resultados da Medida-F em Relação aos Conjuntos de Dados *Jain*, *Flame* e *Spiral* para os Algoritmos K-médias, Espectral/K-médias e Espectral/Aglomerativo.

Resultados da Medida-F			
Conjunto de dados	K-médias	Espectral/K-médias	Espectral/Aglomerativo
Jain	0.8288	0.9446	1
Flame	0.806	0.9825	0.8731
Spiral	0.2695	1	1

A aplicação da métrica referente ao conjunto de dados *Jain* mostrou que o melhor desempenho foi obtido pelo algoritmo aglomerativo, obtendo o valor máximo da medida conforme a Tabela 2. Os resultados obtidos pelos outros dois algoritmos foram razoáveis, ficando com valores acima de 0.8 da medida. Em relação ao conjunto de dados *Flame* nenhum dos algoritmos alcançou 100%, porém o melhor resultado foi obtido pelo algoritmo espectral. Finalmente, no último conjunto de dados tanto o método espectral via k-médias quanto o aglomerativo obtiveram o valor máximo, em contrapartida, o método k-médias ficou com valor abaixo de 0.3, o qual foi o pior resultado obtido.

Pode-se analisar o desempenho dos algoritmos a partir Figura 15, que apresenta cada resultado em forma de gráfico de barras. É possível notar que algoritmo k-médias mostrou-se inferior em relação aos demais nos três conjuntos em estudo. O método espectral tradicional alcançou valores satisfatórios quando utilizado, obtendo o seu menor valor no conjunto de dados *Jain*. Já o método aglomerativo obteve resultados ótimos em dois conjunto de dados, ficando com valor abaixo de 1 apenas no conjunto de dados *Flame*.

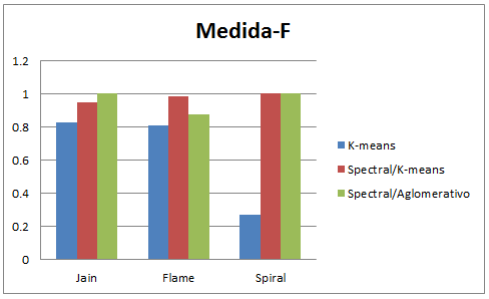


Figura 15: Gráfico dos Resultados de Medida-F para os Algoritmos K-médias, Espectral/K-médias e Espectral/Aglomerativo Utilizados em Experimentos 1

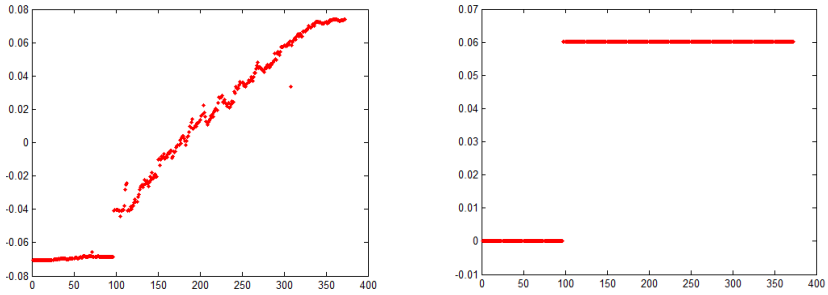
5.3 Discussão

No método espectral o conjunto de dados pode ser visualizado na forma de um grafo, onde cada elemento do conjunto pode ser representado por um vértice e a sua similaridade entre os demais por arestas incidentes nele. Cada valor de uma aresta entre um vértice e outro pode ser calculado por uma função de similaridade, assim, é possível obter um grafo que represente todo o conjunto de dados. Porém, existem alguns métodos que estabelecem como este grafo pode ser construído. O mais geral é o grafo completo, que estabelece uma relação de similaridade entre todos os indivíduos do conjunto gerando um grande número de arestas, o que pode deixar grande a matriz de distâncias a ser calculada. Uma maneira alternativa é construir o grafo utilizando um raio de similaridade entre um elemento em relação ao demais, desse modo, é possível estudar a semelhança entre um conjunto de elementos menor do que o conjunto global. Esta abordagem traz algumas dificuldades de interpretação quando utilizada para representar conjuntos de dados com formato de curva, um exemplo seria o conjunto de dados *Spiral* visto anteriormente. Outra possibilidade a ser considerada é de construir o grafo a partir de um elemento e seus k-vizinhos mais próximos, assim como a abordagem anterior esta possibilidade ajuda a estudar localmente a similaridade em um conjunto menor do que o global.

Quando utilizada a abordagem de construção do grafo usando o método dos k-vizinhos mais próximos ocorre a possibilidade de alternar o número de vizinhos de um elemento. Com esta alternância, o algoritmo espectral pode ser aplicado ao conjunto de dados com a possibilidade de estudar a similaridade localmente ou expandir sua visão dentro do conjunto. Consequentemente, o resultado do agrupamento será sensível a este parâmetro, podendo gerar diferentes grupos para determinados valores de k. Deste modo, o número de k-vizinhos está diretamente relacionado ao conjunto dos k-menores autovetores da matriz Laplaciana, onde ocorre a execução de outro algoritmo de agrupamento, no caso geral, do k-médias.

O que muitas vezes pode gerar um resultado ruim no agrupamento é quando os k-menores autovetores possuem uma curvatura, nesse caso, aplicação do k-médias não é

recomendável. No experimento realizado com o conjunto de dados *Jain* tomou-se os dez vizinhos mais próximos para a construção do grafo de similaridade, gerando como resultado intermediário a representação dos autovetores conforme na Figura 15(a). Nota-se que os autovetores não têm um comportamento linear, assim, o k-médias não irá obter bom resultado quando aplicado neste conjunto. Já na Figura 15(b) pode-se inferir que existem dois grupos bem definidos, isto porque o número de vizinhos mais próximos estabelecido foi de 5. Isto mostra a sensibilidade do método espectral em relação a vizinhança de elementos de um conjunto, ficando a trabalho do usuário acertar uma vizinhança correta para a construção do grafo. Todavia, o método de agrupamento por aglomeração torna-se uma opção para contornar o problema com a geometria dos autovetores, obtendo uma sensibilidade menor em relação a vizinhança do que o método proposto por [1].



(a) 10 vizinhos mais próximos (b) 5 vizinhos mais próximos

Figura 16: K Menores Autovetores do Conjunto de Dados *Jain*

Assim, uma das vantagens do método por aglomeração é a menor sensibilidade ao número de k-vizinhos escolhidos para montar o grafo de similaridade. Conforme o crescimento do número de vizinhos o método por aglomeração também pode variar os resultados do agrupamento, visto que a dependência da geometria do conjunto de dados influencia muito neste contexto.

5.4 Experimentos 2

Nesta segunda seção de experimentos foram escolhidos conjuntos de dados disponíveis em [17]. Cada conjunto foi obtido de acordo com funções utilizadas para criação de dados artificiais no *software* Matlab. Em cada experimento a seguir existe um determinado problema de agrupamento relacionado a sua estrutura. O primeiro conjunto chamado *Two Spirals* possui curvaturas acentuadas que dificultam a aplicação dos algoritmos, além disso, cada elemento (x_i, x_j) possui visivelmente dois vizinhos mais próximos. No contexto de métodos que utilizam *k-nearest neighbour algorithm*, se o usuário informar o número incorreto de k-vizinhos, então a ”visão”do algoritmo no conjunto será maior, podendo causar um agrupamento equivocado.

No segundo conjunto de dados *ClusterinCluster* existem dois grupos formados por duas concentrações circulares de pontos, uma pequena e outra maior. Neste caso, o de-

safio 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.

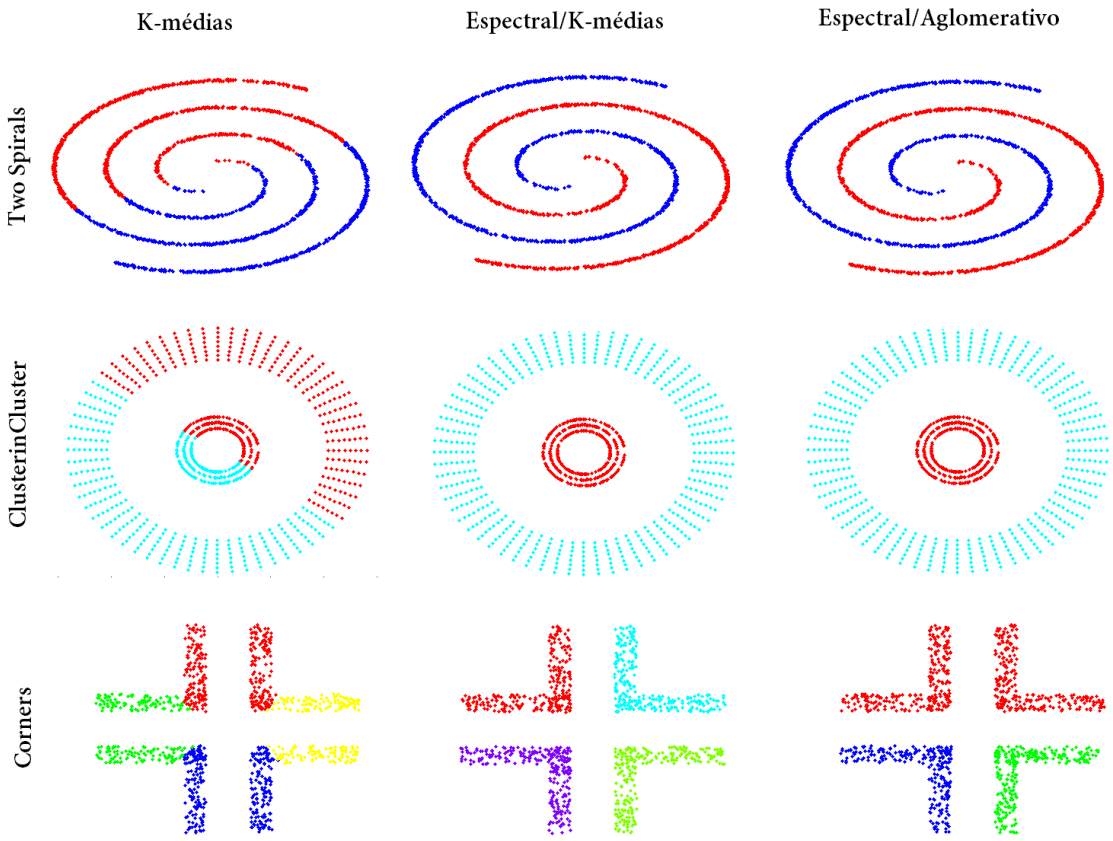


Figura 17: Resultados dos Experimentos Referentes aos Conjuntos de Dados *Two Spirals*, *ClusterinCluster* e *Corners* para os Algoritmos K-médias, Espectral/K-médias e Espectral/Aglomerativo.

A partir da Figura 17 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 dos métodos espectrais em agrupar corretamente os grupos. Já a aplicação do k-médias resultou em uma partição linear no grafo, gerando também dois grupos, porém com resultados não satisfatórios. Neste experimento os valores de entrada informados foram: desvio padrão com $\sigma = 6.0369$, número de grupos $k = 2$ e número

de vizinhos $num\_neighbour = 3$. Neste caso, o parâmetro de maior influência para o bom funcionamento dos algoritmos espectrais é o número de vizinhos que é informado pelo usuário para a construção do grafo de similaridade. Se a vizinhança informada fosse um valor maior do que 10 vizinhos, então o método gera um conjunto de autovetores que não representam um padrão nos dados, sendo assim o agrupamento a ser gerado não irá corresponder a estrutura correta dos grupos.

No experimento referente ao conjunto *ClusterinCluster*, os resultados obtidos pelo k-médias 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. Os parâmetros repassados aos métodos espectrais foram: desvio padrão com $\sigma = 2.5569$, número de grupos $k = 2$ e número de vizinhos $num\_neighbour = 10$. O desvio-padrão, neste caso, pode até ser um fator desconsiderado, pois não representa exatamente os desvios de cada grupo corretamente. Este valor informado é uma média do desvio-padrão dos dois grupos do conjunto, sem a sua utilização, o grafo de similaridade seria construído apenas em função da distância entre os elementos do conjunto. Incluindo o valor de desvio padrão informado os métodos espectrais obtiveram resultados satisfatórios, tanto a abordagem via k-médias quanto via aglomeração geraram os grupos corretos.

No último conjunto de dados *Corners* apenas o método espectral via k-médias obteve corretamente os grupos. Os parâmetros utilizados neste experimento foram: desvio padrão com $\sigma = 5.6805$, número de grupos $k = 4$ e número de vizinhos $num\_neighbour = 10$. Em especial, neste conjunto, todos os parâmetros são relevantes, pois os grupos estão simetricamente posicionados no plano, ou seja, não é dificil obter uma partição incorreta no conjunto. Devido a simetria dos dados, o algoritmo k-médias agrupou os retângulos de acordo com os centroides posicionados ao norte, sul, leste e oeste do conjunto. Já o algoritmo aglomerativo agrupou equivocadamente um dos grupos devido a distribuição dos autovetores no plano bidimensional, ver seção (5.5).

Tabela 3: Resultados da Medida-F em Relação aos Conjuntos de Dados *Two Spirals*, *ClusterinCluster* e *Corners* para os Algoritmos K-médias, Espectral/K-médias e Espectral/Aglomerativo.

Resultados da Medida-F			
Conjunto de dados	K-médias	Espectral/K-médias	Espectral/Aglomerativo
Two Spirals	0.4797	1	1
ClusterinCluster	0.4138	1	1
Corners	0.2813	1	0.6667

A aplicação da medida-F referente ao conjunto *TwoSpirals* mostrou bons resultados em relação aos métodos espectrais conforme Tabela 3. A respeito da performance obtida

pelo k-médias era de se esperar um resultado em torno de 50% de medida-F, visto que, em parte o algoritmo coloca elementos em um grupo correto na mesma proporção em que coloca elementos em um grupo errado. Em relação ao conjunto *ClusterinCluster*, a metodologia de agrupamento espectral obtém 100% de medida-F em ambos algoritmos, porém, de modo similar ao resultado do conjunto anterior o k-médias não mostra eficácia na sua utilização. Os resultados referentes ao conjunto *Corners* torna evidente o resultado ótimo do método espectral via k-médias, respectivamente, seguidos pelos resultados obtidos via aglomeração e k-médias puro. Vale ressaltar que o percentual de acerto de método por aglomeração fica em torno de 75%, já o percentual de erro é de 25%, gerando o valor de medida-F de 0.66.

Figura 18: Gráfico dos Resultados de Medida-F para os Algoritmos K-médias, Espectral/K-médias e Espectral/Aglomerativo Utilizados em Experimentos 2

De modo geral, é possível observar na Figura 18 que nesta seção de experimentos os métodos espectrais obtiveram bons resultados, enfatizando a utilização dos algoritmos em conjuntos que possuam uma determinada curvatura no plano. Como pode-se notar, o caso em que o método por aglomeração não se mostrou 100% eficiente foi em relação ao conjunto *Corners*. Em relação ao k-médias, todos os resultados de medida-F ficaram abaixo de 50%, como pode-se verificar na Figura 18.

5.5 Discussão

Nesta seção, uma atenção especial é dada ao experimento com o conjunto de dados *Corners*. Como foi possível ver, os resultados de agrupamento deste conjunto foram insatisfatórios em relação aos métodos k-médias e espectral aglomerativo. Porém, um fato a ser levado em consideração é de que o método espectral via k-médias nem sempre irá estabelecer o melhor resultado de agrupamento, mesmo neste conjunto em estudo. A explicação para este fato decorre da execução do k-médias, uma vez que este método é utilizado então os centroides são postos de maneira aleatória no conjunto de autovetores, podendo resultar em um agrupamento incorreto. A Figura 5.5 mostra a configuração dos autovetores no plano bidimensional, sendo este o conjunto em que os algoritmos k-médias e aglomerativo são aplicados.

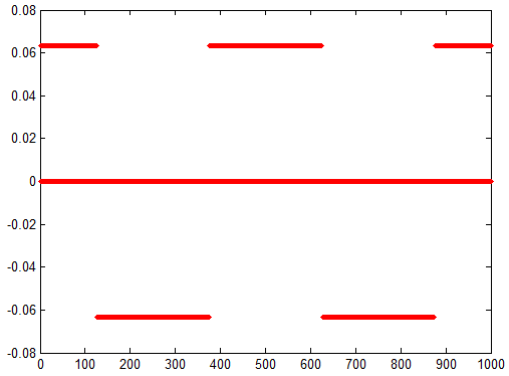


Figura 19: K Menores Autovetores do Conjunto de Dados *Corners*

É fácil ver que na Figura 19 existem cinco concentrações de pontos ao invés de quatro como se gostaria. Ou seja, não é uma tarefa trivial agrupar corretamente os dados considerando este conjunto de autovetores. Na Figura 20 são apresentados mais dois resultados obtidos pelo método espectral via k-médias.

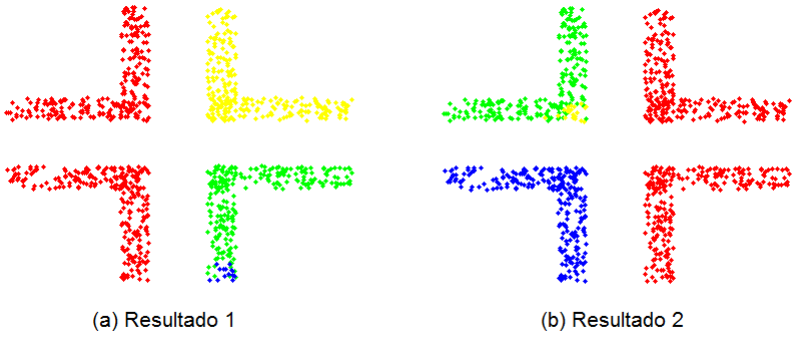


Figura 20: Resultados de Agrupamento Distintos com o Uso do K-médias

A Figura 20 apresenta apenas mais duas execuções distintas do método k-médias, porém estes resultados podem variar ainda mais, dependendo como o método inicializa seus centroides. Na Figura 17 foi apresentado o melhor resultado de agrupamento possível para este conjunto de dados, ou seja, o usuário do algoritmo pode ter que executar o código diversas vezes até obter o resultado que se espera. Logicamente, isto não é o ideal para uma metodologia, principalmente, se utilizada em conjuntos de dados não artificiais, onde não se sabe nada a respeito de possíveis grupos. Já o método aglomerativo, apesar de não ter acertado corretamente todos os grupos, é mais estável. Não há possibilidade de variação nos seus resultados quando utilizado mais de uma vez.

5.6 Experimentos 3

Dando continuidade a utilização dos conjuntos de dados provenientes de [17], nesta seção são descritos mais três experimentos. No primeiro conjunto, chamado *Crescentfull-moon*, existem dois grupos formados por duas meia luas, uma pequena e outra maior, conforme Figura 21. A possibilidade de ocorrer um agrupamento que respeite a distribuição dos elementos no plano depende, neste caso, do uso de algoritmos que não estabelecem grupos apenas de forma linear.

O segundo conjunto de dados *Outlier* possui visualmente quatro grupos, sendo duas concentrações de elementos em formato de meio círculo e outras duas menores concentrações de formato circular. Este conjunto trata de um dos importantes problemas relacionados ao agrupamento, o problema com *outliers*. Um *outlier* é um valor ou um conjunto de valores que diferem do restante de uma distribuição, podendo ter valores considerados muito elevados ou muito pequenos. No caso do conjunto de dados em estudo, os *outliers* fazem com que o conjunto fique ”desbalanceado”, dificultando o agrupamento correto. Neste tipo de problema não é ideal fixar um número k de vizinhos mais próximos, isto porque este valor pode variar em relação a cada grupo.

O último conjunto de dados *Halfkernel* possui uma concentração de pontos em forma de círculo e outra mais abaixo em forma de meia lua conforme Figura 21. Sua principal dificuldade está relacionado a curvatura da meia lua e a proximidade de seus elementos em relação ao círculo na parte superior. É notável que os vários elementos do grupo formado pelo círculo possuem uma distância menor em relação a alguns elementos da meia lua, neste caso, é possível que ocorra um agrupamento dos elementos que estão nas bordas dos dois grupos.

Observando a Figura 21 pode-se concluir que o único algoritmo que agrupou equivocadamento o conjunto de dados *Crescenfullmoon* foi o k-médias. Os parâmetros indicados aos algoritmos foram: desvio padrão $\sigma = 7.6481$, número de grupos $k = 2$ e $num\_neighbour = 10$. Em ambos os métodos espectrais os resultados de agrupamento foram satisfatórios, evidenciando corretamente a geometria do conjunto.

O pior resultado de agrupamento dos métodos em estudo pode-se ver na Figura 21 no conjunto de dados *Outlier*. Em ambos os algoritmos não houve 100% de eficácia, apesar disso, o método que melhor identificou os grupos foi o aglomerativo, que agrupou corretamente 3 subconjuntos. Para a execução dos métodos foram informados: desvio padrão $\sigma = 17.6600$, número de grupos $k = 4$ e $num\_neighbour = 10$.

No último conjunto de dados desta sequência de experimentos, os métodos espectrais obtiveram corretamente os grupos formados pelas duas meia luas do conjunto *Halfkernel*. Com os parâmetros de desvio padrão $\sigma = 11.2024$, número de grupos $k = 2$ e $num\_neighbour = 10$ os autovetores representaram corretamente a estrutura dos dados, sendo possível um bom agrupamento baseado no método espectral. No caso do algo-

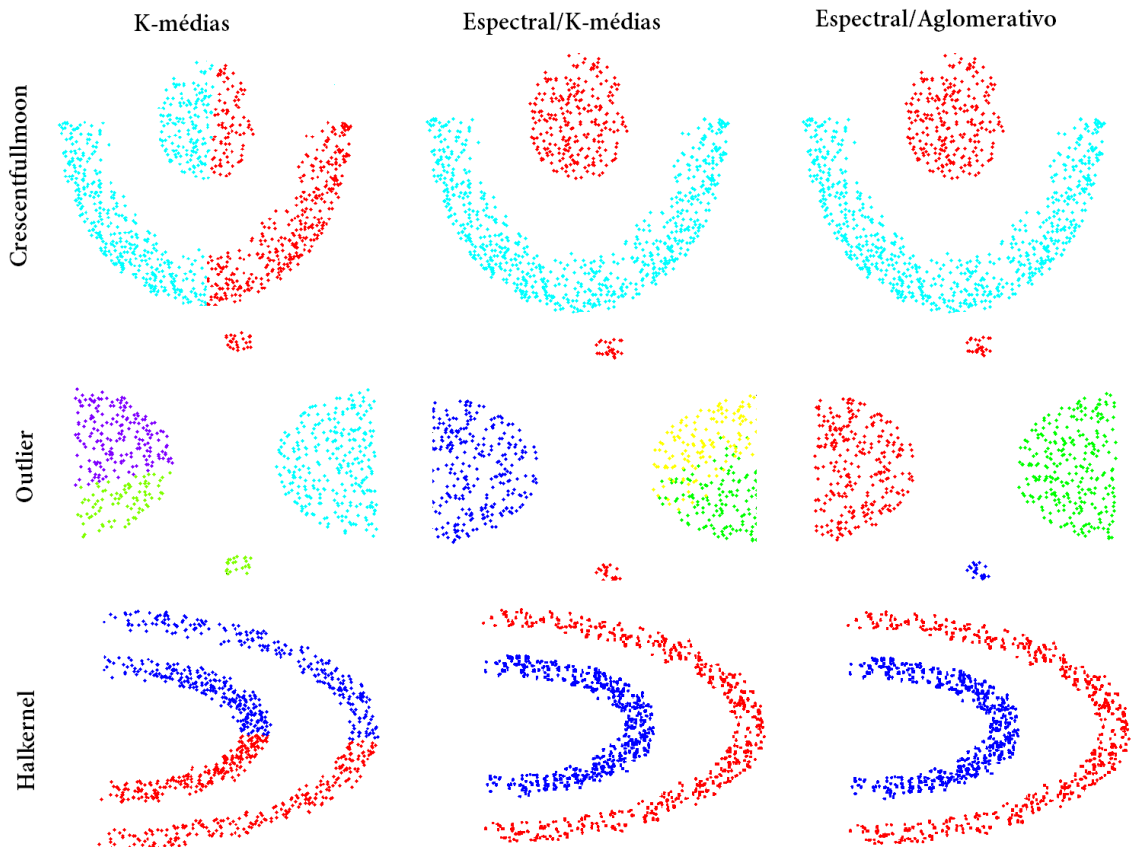


Figura 21: Resultados dos Experimentos Referentes aos Conjuntos de Dados *Crescentfullmoon*, *Outlier* e *Halfkernel* para os Algoritmos K-médias, Espectral/K-médias e Espectral/Aglomerativo.

ritmo k-médias, o resultado de agrupamento gerou dois grupos, de maneira que ambos não correspondem a geometria correta dos dados.

Analisando os resultados da Tabela 4 pode-se concluir que, no geral, o algoritmo espectral aglomerativo obteve melhores resultados de medida-F. Em referência aos três conjuntos de dados, o método por aglomeração obteve seu menor resultado no conjunto *Outlier*, ficando em torno de 95%. Em relação ao método espectral via k-médias os valores de medida-F foram satisfatórios, evidenciando apenas uma porcentagem de erro ao agrupar de maneira equivocada os grupos do conjunto *Outlier*. No que se refere ao algoritmo k-médias, os resultados não representaram que o método foi eficiente na sua utilização nestes experimentos. No geral, os resultados obtidos pelo k-médias foram insatisfatórios.

De acordo com a Figura 22 é possível evidenciar os ótimos resultados do método aglomerativo nos conjuntos *Halfkernel* e *Crescentfullmoon*. Além disso, fica claro que em relação ao conjunto *Outlier*, nenhum método foi totalmente eficaz em descobrir corretamente os grupos deste conjunto. De fato, é um desafio interessante para trabalhar-se com este conjunto de dados e utilizar uma metodologia que seja capaz de inferir na obtenção

Tabela 4: Resultados da Medida-F em Relação aos Conjuntos de Dados *Crescentfullmoon*, *Outlier* e *Halfkernel* para os Algoritmos K-médias, Espectral/K-médias e Espectral/Aglomerativo.

Resultados da Medida-F			
Conjunto de dados	K-médias	Espectral/K-médias	Espectral/Aglomerativo
Crescentfullmoon	0.4138	1	1
Outlier	0.6309	0.772	0.9583
Halfkernel	0.4026	1	1

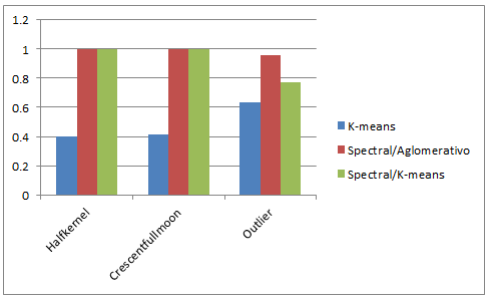


Figura 22: Gráfico dos Resultados de Medida-F para os Algoritmos K-médias, Espectral/K-médias e Espectral/Aglomerativo Utilizados em Experimentos 3

correta dos grupos.

5.7 Experimentos 4

Nesta 4ª 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 a proximidade entre os elementos dos dois grupos o fator distância não torna-se suficiente para a determinação correta do agrupamento.

Analisando a Figura 23 nota-se, primeiramente, que o resultado de agrupamento pelo método k-médias não obtém corretamente os grupos formados pelos dois anéis. Esta configuração do conjunto não permite que as posições dos centroides se ajustem de acordo

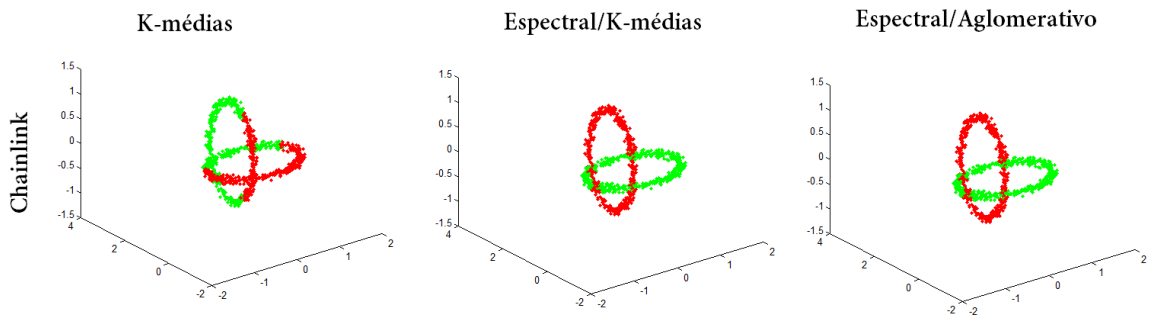


Figura 23: Resultados dos Experimentos Referente ao Conjunto de Dados *Chainlink* para os Algoritmos K-médias, Espectral/K-médias e Espectral/Aglomerativo.

com a curvatura dos dados, desse modo, ocorre uma partição linear do conjunto gerando dois grupos que não correspondem a geometria correta dos dados. Os parâmetros utilizados pelos algoritmos foram definidos de acordo com: desvio padrão $\sigma = 0.6768$, número de grupos $k = 2$ e $num\_neighbour = 5$. Para os métodos de particionamento espectral, a criação do grafo de similaridade considerando os 5 vizinhos mais próximos auxilia diretamente no resultado de agrupamento obtido pelos autovetores da matriz Laplaciana. Se o grafo de similaridade fosse construído com um número de vizinhança muito grande, a geometria dos autovetores seria muito aleatória, não representando de fato o formato do conjunto original. Sendo assim, os métodos espectrais obtiveram um resultado de agrupamento satisfatório considerando um número de vizinhança pequeno.

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 24, 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.

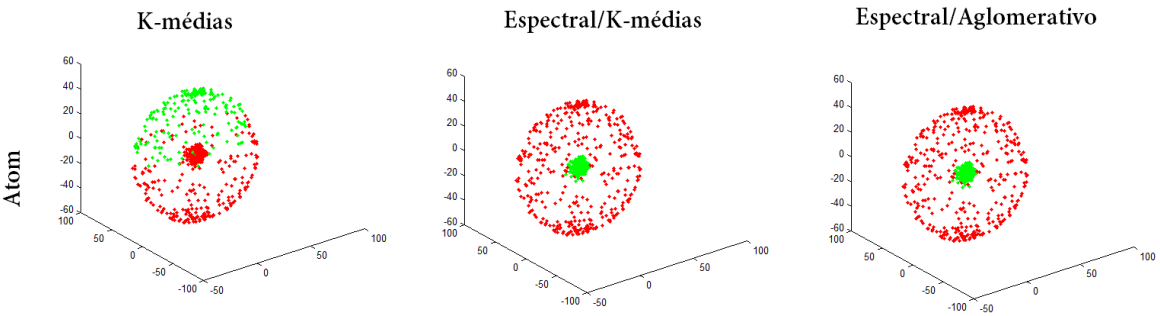


Figura 24: Resultados dos Experimentos Referente ao Conjunto de Dados *Atom* para os Algoritmos K-médias, Espectral/K-médias e Espectral/Aglomerativo.

Considerando os parâmetros de desvio padrão $\sigma = 20.5545$, número de grupos $k = 2$ e $num\_neighbour = 10$ informados aos algoritmos, os resultados de agrupamento são apresentados na Figura 24. Como pode-se notar o resultado de agrupamento dos métodos espectrais foram ótimos, permitindo a descoberta correta dos dois grupos do conjunto *Atom*. Em relação ao algoritmo k-médias, o agrupamento novamente ocorreu de forma equivocada, não evidenciando as duas concentrações esféricas de dados.

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.

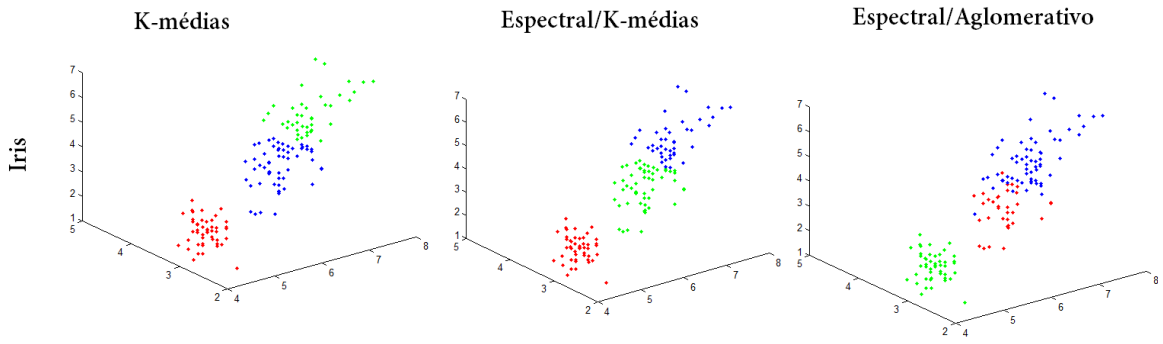


Figura 25: Resultados dos Experimentos Referente ao Conjunto de Dados *Iris* para os Algoritmos K-médias, Espectral/K-médias e Espectral/Aglomerativo.

Para executar os algoritmos de agrupamento foram informados os seguintes parâmetros: desvio padrão $\sigma = 1.9755$, número de grupos $k = 3$ e $num\_neighbour = 5$. Na Figura 25 nota-se que os algoritmos k-médias e espectral via k-medias apresentam resultados bem similares. Já o método espectral aglomerativo difere dos outros dois, ressaltando elementos que não pertencem diretamente a uma determinada concentração. Também pode-se notar que grupo mais abaixo do conjunto é o mesmo em ambos os resultados de agrupamento dos algoritmos, neste caso, este grupo é referente a *Iris setosa*. Desse modo, os três métodos diferem apenas na parte superior do conjunto *Iris*, onde concetram-se os grupos *Iris virginica* e as *Iris versicolor*, os quais apresentam uma maior concentração de pontos e proximidade entre si.

A partir da análise da Tabela 5 pode-se concluir que nos dois primeiros conjuntos de dados os algoritmos espectrais foram 100% eficazes, evidenciando corretamente a configuração dos grupos no espaço tridimensional. Um fato interessante nos resultados

Tabela 5: Resultados da Medida-F em Relação aos Conjuntos de Dados *Chainlink*, *Atom* e *Iris* para os Algoritmos K-médias, Espectral/K-médias e Espectral/Aglomerativo.

Resultados da Medida-F			
Conjunto de dados	K-médias	Espectral/K-médias	Espectral/Aglomerativo
Chainlink	0.348	1	1
Atom	0.6146	1	1
Iris	0.8085	0.8148	0.8718

do conjunto *Iris* foi o melhor desempenho do algoritmo k-médias em relação ao método de particionamento espectral tradicional. Porém, o método espectral por aglomeração mostrou-se melhor em relação aos outros dois, obtendo uma resultado em torno de 87% de medida-F, além disso, melhorando cerca de 6% o resultado em relação ao método espectral via k-médias.

6 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

Neste trabalho foram apresentados conceitos básicos de particionamento espectral de grafos, juntamente com a apresentação do principal algoritmo de agrupamento utilizado para agrupamento espectral de dados. Foi possível notar que o algoritmo k-médias nem sempre é uma boa alternativa a ser usado no agrupamento dos autovetores. A sua sensibilidade a parâmetros de inicialização e de vizinhança são fatores que impulsionaram o desenvolvimento desta pesquisa.

Sendo assim, a principal contribuição foi inserir ao método de agrupamento espectral um algoritmo baseado em aglomeração para agrupar os k-menores autovetores da matriz Laplaciana. Como pode ser visto no capítulo de experimentos, o uso do método aglomerativo obteve bons resultados, superando sempre o algoritmo k-médias, e em alguns casos, o algoritmo espectral via k-médias. As medidas de desempenho obtidas via medida-F mostraram que, em uma visão geral, o método por aglomeração obteve resultados coerentes ficando sempre acima de 87% de acerto, excluindo o caso do conjunto de dados *Corners* com 66% de acerto.

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-médias. Para o método espectral aglomerativo seria útil o estudo de outras métricas e analisar suas implicações.

Referente ao método proposto, este ainda pode ser melhorado, principalmente, em termos de complexidade computacional, assim possibilitaria a sua utilização em bancos de dados de grande volume. A etapa de maior custo para a metologia refere-se ao cálculo dos autovetores da matriz Laplaciana, que geralmente, é calculado via *solver* dentro do *software* utilizado. Isto é um problema que pode ser contornado, mas seria uma contribuição à parte deste trabalho.

Além disso, para trabalhos futuros seria útil investigar a relação entre a geometria dos

autovetores e os parâmetros utilizados para montar o grafo de similaridade. Conforme foi visto na seção (5.3), na medida em que se varia o número de vizinhos para a construção do grafo, também varia a distribuição dos autovetores no plano. Um caso onde não ocorre esta variação ocorre quando utiliza-se o grafo de similaridade completo, porém, computacionalmente o custo para execução desta tarefa pode ser alta, e ainda, o método espectral teria apenas uma visão global do conjunto de dados.

Por fim, este trabalho apresentou um algoritmo de agrupamento espectral baseado em aglomeração que pode ser utilizado como uma alternativa ao algoritmo espectral tradicional. Enfatizando os bons resultados, conclui-se que a metodolgia é promissora e pode ser aprimorada em trabalhos futuros.

REFERÊNCIAS

[1] M. Jordan A. Ng and Y.Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*. 2001.

[2] H. Chang and D.Y. Yeung. Robust path-based spectral clustering. In *Pattern Recognition*, pages 191–203. 2008.

[3] X. Chen and D. Cai. Large scale spectral clustering with landmark-based representation. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*. 2011.

[4] L. F. Mendonca D. F. dos Santos and M. G. Teixeira. Um algoritmo de agrupamento heterogeneo para formacao de grupos de aprendizagem. In *Proceeding Series of the Brazilian Society of Applied and Computational Mathematics*. 2015.

[5] L. Huang D. Yan and M.I. Jordan. Fast approximate spectral clustering. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 907–916. 2009.

[6] L. Fu and E. Medico. Flame, a novel fuzzy clustering method for the analysis of dna microarray data. In *BMC bioinformatics*. 2007.

[7] A. Jain and M. Law. Data clustering: A user's dilemma. In *Lecture Notes in Computer Science*, pages 1–10. 2005.

[8] W. Kot lowski K. Dembczynski, A. Jachnik and W. Waegeman. Optimizing the f-measure in multi-label classifcation: Plug-in rule approach versus structured loss minimization. In *International Conference on Machine Learning*, pages 1130–1138. 2013.

[9] M. Steinbach L. Ertoz and V. Kumar. Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In *SDM*, pages 47–58. 2003.

[10] K. Ramamohanarao L. Wang, C. Leckie and J. Bezdek. Approximate spectral clustering. In *Advances in Knowledge Discovery and Data Mining*, pages 134–146. 2009.

[11] V. Luxburg. A tutorial on spectral clustering. In *Statistics and computing*, pages 395–416. 2007.

[12] L Karlsson M Langkvist and A Loutfi. A review of unsupervised feature learning and deep learning for time-series modeling. In *Pattern Recognition Letters*, pages 11–24. 2014.

[13] L. Z. Manor and P. Perona. Self-tunning spectral clustering. In *Advances in Neural Information Processing Systems (NIPS17)*. 2004.

[14] V. Trevisan N. Abreu, R. Del-Vecchio and C.T. Vinagre. Teoria espectral de grafos-uma introducao. 2013.

[15] R. Gribonval N. Tremblay, G. Puy and P. Vandergheynst. Compressive spectral clustering. In *arXiv preprint arXiv:1602.02018*. 2016.

[16] B. Nadler and M. Galun. Fundamental limitations of spectral clustering. In *Advances in neural information processing systems*, pages 1017–1024. 2006.

[17] S. Overflow. http://stackoverflow.com/questions/16146599/create-artificial-data-in-matlab. 2017.

[18] Q. Zhu P. Yang and B. Huang. Spectral clustering with density sensitive similarity function. In *Knowledge-Based Systems*, pages 621–628. 2011.

[19] W. Pentney and M. Meila. Spectral clustering of biological sequence data. In *Association for the Advancement of Artificial Intelligence*, pages 845–850. 2005.

[20] M. Steinbach P.N. Tan and V. Kumar. Introduction to data mining. In *Ciencia Moderna*. 2009.

[21] M. Saerens. The principal components analysis of a graph, and its relationships to spectral clustering. In *European Conference on Machine Learning*, pages 371–383. 2004.

[22] Tomoya Sakai and Imiya. Fast spectral clustering with random projection and sampling. In *Machine Learning and Data Mining in Pattern Recognition*, pages 372–384. 2009.

[23] S. Tsironis and M. Sozio. Accurate spectral clustering for community detection in mapreduce. In *Neural Information Processing Systems*. 2013.

[24] H. Bai C. Lin Y. Song, W. Chen and E. Chang. Parallel spectral clustering. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. 2008.