UNIVERSIDADE FEDERAL DE RIO GRANDE

PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL

Bruna Souza dos Santos

Rio Grande, Setembro de 2017.

UNIVERSIDADE FEDERAL DE RIO GRANDE
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL

Bruna Souza dos Santos

Agrupamento de dados utilizando espectro de grafos

Dissertação de Mestrado apresentado ao Programa de Pós Graduação em Modelagem Computacional da Universidade Federal do Rio Grande. Orientação: Profª. Drª. Catia Maria dos Santos Machado.Co-orientação: Profº. Drº. Leonardo Emmendorfer.

Rio Grande, Setembro de 2017

AGRADECIMENTO

Agradeço primeiramente a Deus por ter me fortalecido nos momentos difíceis e me guiado pelo melhor caminho.

Aos meus pais Airton, Ana e Norma pelo apoio, educação e cuidados de sempre. Principalmente por não medirem esforços para eu concluir mais essa etapa em minha vida, través deles agradeço a todos os meus familiares, as palavras de incentivo de cada um foram fundamentais nessa jornada.

Ao meu esposo, pelo companheirismo, nas viagens até Rio Grande, pela paciência nos dias de estudo e por acreditar no meu potencial.

A Bispa Joelize e ao Pastor Alcides pelas orações e através deles agradeço a todos da família Mensageiros de Cristo.

As minhas funcionárias por sempre serem prestativas e cuidarem da loja nos dias de estudo.

Aos professores, Marilton e Viviane, pelas sugestões e contribuições, enriquecendo e tornado o trabalho muito melhor.

A minha orientadora Catia, ao co-orientador Leonardo e ao colega Luciano, obrigada por me orientar, ensinar, compreender e por terem me incentivado até o final vocês foram excepcionais.

RESUMO

O presente trabalho tem como objetivo principal verificar os resultados obtidos, por um recente algoritmo de aglomeração, para o problema de agrupamento de dados. O algoritmo de aglomeração, a partir dos k-menores autovetores da matriz Laplaciana, agrupa um conjunto de dados a partir do grafo de similaridade. Utilizando a comparação com os algoritmos k-médias e espectral via k-médias, sobre um banco de dados da literatura, é possível mostrar que o algoritmo de aglomeração é uma opção promissora no estudo de agrupamentos de dados.

Palavras-chave: Agrupamento, Espectro e Particionamento.

ABSTRACT

The main objective of this work is to verify the results obtained by a recent agglomeration algorithm for the data grouping problem. The algorithm of agglomeration from the k-minor eigenvectors of the Laplacian matrix groups a set of data from the similarity graph. Using the comparison with the k-medium and k-medium spectral algorithms on a literature database, it is possible to show that the agglomeration algorithm is a promising option in the study of data groupings.

Keywords: Grouping, Spectrum and Partitioning.

LISTA DE FIGURAS

LISTA DE EQUAÇÕES

LISTA DE TABELAS

Sumário

CAPÍTULO 1.

1. INTRODUÇÃO

Os problemas de agrupamento surgiram da necessidade de agrupar dados a fim de entender um objeto ou um fenômeno ainda desconhecidos. Na mineração de dados o agrupamento é uma tarefa importante, tendo como objetivo segmentar uma base de dados em grupos de objetos baseando-se na similaridade ou dissimilaridade entre os mesmos. Devido à natureza não supervisionada da tarefa, a busca por uma solução de boa qualidade pode tornar-se um processo complexo (VILMAR, T. *et al*., 2013). A mineração de dados usa análise matemática para derivar padrões e tendências que existem nos dados.

Segundo (CASSIANO, K. M., 2014) a grande vantagem do uso das técnicas de Clusterização é que, ao agrupar dados similares, pode-se descrever de forma mais eficiente e eficaz as características peculiares de cada um dos grupos identificados. Isso fornece um maior entendimento do conjunto de dados original, além de possibilitar o desenvolvimento de esquemas de classificação para novos dados e descobrir correlações interessantes entre os atributos dos dados que não seriam facilmente visualizadas sem o emprego de tais técnicas. Alternativamente, Clusterização pode ser usada como uma etapa de pré-processamento para outros algoritmos, tais como caracterização e classificação, que trabalhariam nos clusters identificados.

A Clusterização pode ser aplicada quando objetivo é reduzir o número de objetos, para um número de subgrupos característicos, levando as observações a serem consideradas como membros de um grupo e reorganizadas segundo características gerais que distinguem os rótulos destes grupos, ou também quando o pesquisador deseja formular hipóteses sobre a natureza dos dados ou examinar hipóteses pré-estabelecidas.

Há muitos algoritmos de agrupamento na literatura (VILMAR, T. *et al*., 2013) que ocasionam uma difícil categorização dos métodos e das abordagens de agrupamento existentes. O trabalho proposto visa principalmente o estudo do comportamento do algoritmo recentemente desenvolvido por (GARIN, G. L,2017) quando comparado com dois algoritmos da literatura k-médias e espectral via k-médias, através da medida-F de similaridade.

1.2. OBJETIVOS

1.2.1. OBJETIVO GERAL

O presente trabalho tem por objetivo geral comparar o algoritmo Aglomerado, recentemente desenvolvido , com os algoritmos k-médias e espectral via k-médias. Para alcançar o objetivo geral os seguintes objetivos específicos são considerados.

1.2.2. OBJETIVOS ESPECÍFICOS

- implementar os algoritmos de agrupamentos utilizando o Software livre Octave;
- testar o algoritmo, que utiliza a teoria espectral de grafos na resolução de problemas de agrupamento, sobre um banco dados encontrados na literatura;
- comparar os resultados obtidos com a abordagem baseada na distância entre pontos (distância euclidiana);
- comparar os três algoritmos de agrupamento utilizando a medida-F.

1.3. IMPORTÂNCIA DO TRABALHO

O estudo do trabalho é justificado, pelo fato de mostrar que um algoritmo recentemente desenvolvido por (GARIN, G. L,2017) (método baseado em agrupamento espectral) pode ser capaz de identificar padrões em estrutura de grafos onde estão representados os dados. Além disso, é possível mostrar que esse novo algoritmo é competitivo quando comparado com dois algoritmos, um habitualmente utilizado (k-médias) e espectral via k-médias.

O estudo da teoria espectral de grafos na resolução de problemas de otimização combinatória é um tema de pesquisa interessante, pois propriedades matemáticas quando devidamente aproveitadas contribuem para o desenvolvimento de algoritmos.

1.4. ESTRUTURA DO TRABALHO

O trabalho está organizado da seguinte forma: O capítulo 1 é dedicado à introdução, objetivos e importância do trabalho. O capítulo 2, apresenta a fundamentação teórica sobre agrupamento, particionamento, algoritmos de agrupamento e medida de qualidade de agrupamento. O capítulo 3, apresenta a metodologia fundamentada na Teoria Espectro e Particionamento em grafos. O capítulo 4, apresenta os testes computacionais, resultados, discussão e considerações finais.

CAPÍTULO 2.

Neste capítulo,na Seção 2.1,é apresentada uma breve revisão bibliográfica sobre formas de agrupamentos e algoritmos de agrupamentos encontrados na literatura. A Seção 2.2, traz uma breve explicação sobre algumas medidas de qualidade. A Seção 2.3, destaca alguns trabalhos relacionados com a tarefa de agrupamento de dados convergindo com a proposta do trabalho.

2. REVISÃO BIBLIOGRÁFICA

2.1 AGRUPAMENTOS

A mineração de dados pode ser considerada como uma parte do processo de Descoberta de Conhecimento em Banco de Dados (KDD – Knowledge Discovery in Databases). Combina ferramentas de diferentes áreas, como aprendizagem de máquina, estatística, banco de dados, sistemas especialistas e visualização de dados (ANDA, 1999). Pesquisadores relatam que esse processo pode ser dividido em algumas tarefas, dentre elas, a tarefa de classificação e a tarefa de agrupamento.

A tarefa de agrupamento (clustering) é altamente popular na área de mineração de dados, muitas vezes utilizada como um passo inicial na análise exploratória de conjuntos de dados complexos. É intensamente estudada devido à sua aplicabilidade em diversas áreas de conhecimento (e.g., marketing, engenharia, medicina) e é apresentada como uma abordagem não-supervisionada, pois não se sabe a priori a que grupo cada objeto pertence. (GUEDES, G. P. *et al*., 2016).

Agrupamentos podem ser classificados como sendo do tipo particional (divisão dos objetos dentro do conjunto de objetos) e hierárquico (organizados em forma de árvores). É possível classificar um agrupamento como sendo do tipo particional quando, realizando uma simples divisão dos objetos dentro do conjunto de objetos, cada objeto permaneça em um único grupo. Ao passo que, se o objetivo for que os grupos tenham subgrupos, então o agrupamento é do tipo hierárquico. Os agrupamentos hierárquicos são organizados em forma de árvores. Ocasionalmente, as folhas destas árvores são formadas por grupos de um único objeto (TAN, P-N; *et al.*,2009).

Existem, outras formas de classificar um agrupamento. Um agrupamento pode ser exclusivo, interseccionado ou difuso. Agrupamentos exclusivos são agrupamentos

onde cada objeto está presente exclusivamente em um único grupo. Agrupamentos interseccionados, quando um mesmo objeto faz parte de mais de um grupo ao mesmo tempo. Agrupamentos difuso, quando cada objeto pertencente a um grupo possui um peso por ser membro do grupo (o peso tem valor que varia entre 0 e 1).

Além das formas de classificação dos agrupamento já citadas, um agrupamento pode também ser completo ou parcial. O agrupamento completo atribui todos objetos do conjunto para algum grupo. O agrupamento parcial é aquele em que pode existir objetos do conjunto que não estejam em nenhum grupo ao final do agrupamento (TAN, P-N; *et al.*, 2009). As Figuras 1, 2, 3 e 4 apresentam algumas formas de classificação de agrupamentos.

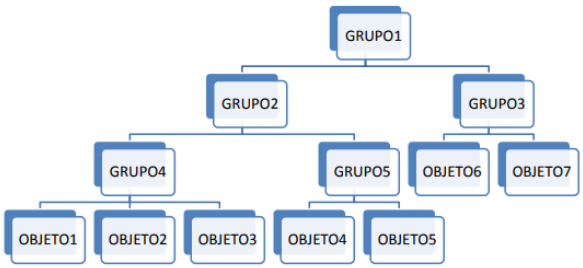


Figura 1: Agrupamento Hierárquico
Fonte: PEREIRA, I. A. 2013.

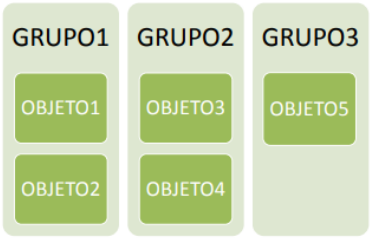


Figura2: Agrupamento Particional Exclusivo.
Fonte: PEREIRA, I. A. 2013.

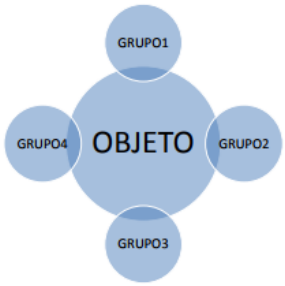


Figura 3: Agrupamento Interseccionado
Fonte: PEREIRA, I. A. 2013.

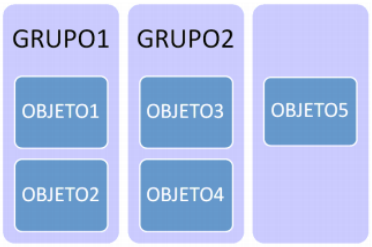


Figura4: Agrupamento Parcia
Fonte: PEREIRA, I. A. 2013.

Segundo (ROCHA, F. C., 2015) as principais abordagens utilizadas para agrupamentos de dados são as abordagens baseadas em protótipos e baseadas em densidades. Na medida que for possível determinar protótipos a partir de um conjunto de objetos, esses objetos reconhecidos como protótipos poderão agrupar outros objetos do conjunto mediante a existência de alguma semelhança entre eles (entre o protótipo e os outros objetos do conjunto). A determinação dos protótipos deve ser realizada a partir do cálculo do centroide ou da média do conjunto. No entanto, há conjunto de

objetos em que a determinação dos protótipos torna-se impraticável, em virtude da necessidade do cálculo da média (conjuntos de objetos compostos por dados categóricos) (Luz, D., 2003). Para estes conjuntos é preciso mudar a abordagem e, em vez de determinar os protótipos, determinar zonas (ou regiões) de alta densidade para realizar o agrupamento. Assim, grupos são formados a partir da divisão entre zonas de alta e de baixa densidade (ROCHA, F. C., 2015).

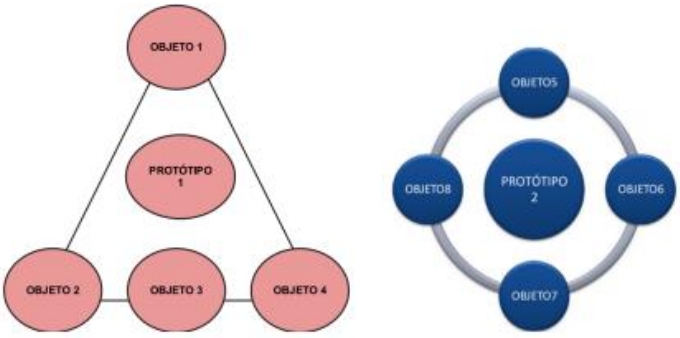


Figura 5:Abordagem de agrupamento Baseada em Protótipo
Fonte: PEREIRA, I. A. 2013.

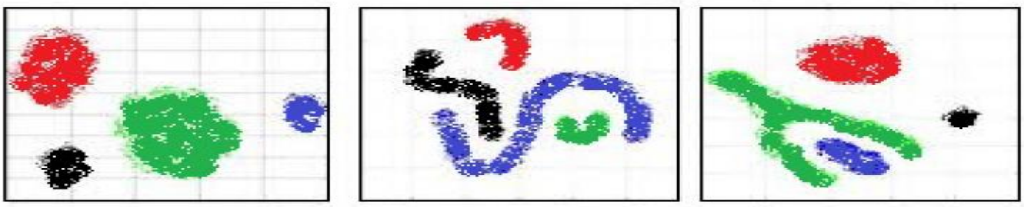


Figura 6:Abordagem de agrupamento por densidade.
Fonte: PEREIRA, I. A. 2013.

Muitos algoritmos, desenvolvidos com a tarefa de agrupar dados, podem ser considerados como procedimentos orientados por uma função objetivo. Tipicamente, o espaço de busca é bastante grande, posto que cada estado desse espaço corresponde a uma possível partição do conjunto de objetos. O procedimento de otimização procura encontrar uma partição na qual os objetos de cada grupo sejam semelhantes e objetos em grupos distintos sejam dissimilares (HAN *et al*., 2011).

(GAN *et al*., 2007b) classificam os algoritmos de agrupamentos em duas categorias amplas: algoritmos de agrupamento hierárquico (hierarchicalclustering) e algoritmos de agrupamento particional (partitionalclustering).

Os algoritmos hierárquicos são subdivididos em duas abordagens: aglomerativos (bottom-up) e divisivos (top-down). Esses algoritmos têm seu funcionamento bastante

semelhante. Os algoritmos divisivos consideram que, inicialmente, todo o conjunto de dados está em um grupo e, a cada iteração, particiona esse grupo em grupos menores.

Os algoritmos hierárquicos possuem algumas limitações, dentre elas: i) uma vez que a decisão de combinar dois grupos tenha sido tomada, não pode ser desfeita. ii) nenhuma função objetivo é minimizada diretamente. Além disso, podem ser utilizados apenas em conjuntos de dados relativamente pequenos (RAJARAMAN, A.; ULLMAN, J.D., 2011).

Os algoritmos de agrupamento particional dividem os dados diretamente em um número de grupos, não havendo uma estrutura hierárquica entre os grupos. Assim, dado um número *n* de objetos, os dados são divididos em *k* grupos, considerando que cada objeto se encontre em exatamente em um grupo *k*(GUEDES, G. P. *et al*., 2016).

2.2. MEDIDAS DE QUALIDADE

Segundo, (JAIN, A.K.; DUBES, R.C., 1988) e (KUNCHEVA, L. I., 2004) no contexto do aprendizado supervisionado, há uma variedade grande de medidas para avaliar o modelo gerado: precisão, conferência, recordação (recall) e outros. Dessa forma, é preciso avaliá-los para comparar diferentes algoritmos de agrupamento a fim de evitar a descoberta de padrões em ruído. Comparar duas partições é comparar dois grupos (cluster) e determinar a tendência de grupo (clusteringtendency) de um conjunto de dados. A importância dessa avaliação se faz necessária para identificar se uma estrutura não-aleatória de fato existe nos dados, pois a maioria dos agrupamentos encontram grupos mesmo quando os dados são aleatórios.

As medidas numéricas utilizadas na validação de agrupamento são divididas em três grupos: Índices externos, índices internos e índices relativos. Os índices externos são utilizados para **avaliar** o agrupamento gerado de acordo com a estrutura pré-especificada, imposta ao conjunto de dados (índice Rand ajustado e índice de Jaccard). Os índices internos são usados para **medir** a qualidade de um agrupamento baseado apenas nos dados originais (instâncias ou matrizes de similaridade), alguns dos índices utilizados são: Índice Davies-Bouldin, Índice Dunn, Silhuetas, medida-F entre outros. Os índices relativos são empregados para **comparar** diversos agrupamentos e decidir qual deles é melhor em algum aspecto. Em geral, pode ser utilizado qualquer um dos índices acima definidos. O índice deve fazer sentido intuitivamente, deve ter uma base teórica e deve ser prontamente computável.

2.2.1.ÍNDICES DAVIES-BOULDIN

O índice Davies-Bouldin (DAVIES, D. L., BOULDINB, D. W., 1979) não depende do número de agrupamentos e do método de partição dos dados, o que o torna adequado para avaliação de algoritmos de partição de dados. O índice é dado em função da razão entre a soma da dispersão interna dos agrupamentos e a distância (separação) entre os agrupamentos. A dispersão interna de um agrupamento i é calculada como segue:

$$S_{i,q} = \left(\frac{1}{|C_i|}\right) \sum_{x \in C_i} \{|x - C_i|^q\}^{\frac{1}{q}} \quad (1)$$

e a distância entre o agrupamento C_i e o agrupamento C_j é definida como:

$$d_{ij,t} = \left\| z_i - z_j \right\|_t \quad (2)$$

onde, $S_{i,q}$ é a raiz q-ésima do q-ésimo momento dos $|C_i|$ pontos no agrupamento C_i em torno da sua média z_i . Se $q = 1$,$S_{i,1}$ é a distância euclidiana média dos vetores no agrupamento i em relação ao centroide deste grupo. Se $q = 2$,$S_{i,2}$ é o desvio padrão da distância dos vetores em relação ao vetor centroide do grupo. Denomina-se $d_{ij,t}$ a distância de Minkowski de ordem t entre os centróides z_i e z_j que caracterizam os agrupamentos C_i e C_j. Subsequentemente calcula-se:

$$R_{i,qt} = max_{j, j \neq i} \left\{ \frac{S_{i,q} + S_{j,q}}{d_{ij,t}} \right\} \quad (3)$$

O índice Davies-Bouldin é então definido como:

$$\frac{1}{C} \sum_{i=1}^{C} R_{i,qt} \quad (4)$$

2.2.2. SILHOUETTES

O coeficiente $Silhouette$ baseia-se na ideia de quanto um objeto é similar aos demais membros do seu grupo, e de quanto este mesmo objeto é distante de outro grupo. Assim, essa medida combina as medidas de coesão e acoplamento

$$S = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (5)$$

Onde $a\ (i)$ é a distância média entre o $i - ésimo$ elemento do grupo e os outros do

$$\bar{S} = \frac{1}{N}\sum_{i=1}^{N} S \quad (6)$$

mesmo grupo. O $b(i)$ é o valor mínimo de distância entre o $i-ésimo$ elemento do grupo e qualquer outro grupo, que não contém elemento, e max é maior distância entre $a(i)$ e $b(i)$. Assim essa, medida combina as medidas de coesão e acoplamento

$$\bar{s} = \frac{1}{N}\sum_{i=1}^{N} S \quad (7)$$

O Coeficiente $Sillhouette$ de um grupo é a medida aritmética dos coeficientes calculados para cada elemento pertencente ao grupo o valor de S situa-se na faixa de 0 a 1.

2.2.3. MEDIDA-F

Segundo (MANNING, C.D. *et al* 2008), o recall (revocação) mede a proporção de objetos corretamente alocados a um agrupamento, em relação total de objetos da classe associada a este agrupamento. A precision (precisão) mede proporção de objetos corretamente alocados a um agrupamento, em relação ao total de objetos deste agrupamento. Assim, a medida F, é a medida harmônica entre o precision e o recall e assume valores que estão no intervalo de [0,1]. O valor zero indica que nenhum objeto foi agrupado corretamente, o valor um, que todos os objetos estão corretamente agrupados. Assim, um agrupamento ideal deve retornar um valor igual a um.

O processo de aprendizado é um processo de escolher uma função apropriada $h(x)$, de um conjunto de funções $(h_1(x), \dots, h_m(x))$. Dessa forma, dada uma predição $h(x) = h_1(x), \dots, h_m(x) \in Y$de um vetor de rótulo binário $y = (y_1, \dots, y_m)$, a medida-F é definida como:

$$F\big(y, h(x)\big) = \frac{(1+\beta^2)\sum_{i=1}^{m} y_i h_i(x)}{\beta^2 \sum_{i=1}^{m} y_i + \sum_{i=1}^{m} h_i(x)} \in [0,1], \quad (8)$$

Mais detalhes sobre o processo de aprendizado supervisionado e medida F podem ser encontrados nos trabalhos de (RUFFINO, L.P. 2011) e (WOJCIECH, K.;*et al.*2013) respectivamente.

2.3. TRABALHOS RELACIONADOS

Muitos trabalhos estão relacionados com a tarefa de agrupamento de dados, ao proceder a revisão bibliográfica foram destacados os seguintes trabalhos:

(STEINLEY, D.; BRUSCO, M. J., 2007), relata que o algoritmo k-means utilizado para particionamento de dados é sem dúvida o algoritmo mais popular. No entanto, o

algoritmo apresenta problemas por obter soluções localmente ótimas e a partição final depender da configuração inicial, fazendo com que a escolha das partições iniciais seja ainda mais importante. O trabalho, avalia procedimentos propostos na literatura e fornece recomendações para melhorar as práticas do algoritmo mais popular para particionamento dados, o algoritmo k-means.

(VON LUXBURG, U., 2007),descreve diferentes matrizes Laplacianas e suas propriedades básicas,apresentando abordagens diferentes do algoritmo de agrupamento espectral. Além disso,mostra vantagens e desvantagens dos diferentes algoritmos de agrupamento espectral.

(SCHAEFFER, S. E., 2007), mostram as definições e os métodos para o agrupamento de grafos, ou seja, encontram um conjuntos de vértices "relacionados" a grafos. Revisam as muitas definições para o que é um cluster em um grafo e medidas de qualidade de cluster. Apresentam algoritmos globais para produzir um cluster para todo o conjunto de vértices de um grafo de entrada. Ainda discutem a tarefa de identificar um cluster para um vértice "semente" pela computação local. Fazem uma abordagem a problemática da avaliação de agrupamentos e benchmarking de algoritmos de cluster.

(KURSUN, O., 2010), o autor utiliza um método de agrupamento espectral, o método de estimação de densidades através de um núcleo (Kernel) para agrupar objetos de dados utilizando os autovetores derivados dos dados. Eles propõe uma análise a sensibilidade em grupos com valores fora do padrão e determina a banda do Kernel a partir dos dados.

(XU, R.; WUNSCH, D., I., 2005),os autores pesquisaram algoritmos de agrupamento para conjuntos de dados que aparecem em estatísticas, informática e aprendizagem de máquinas e ilustram suas aplicações em alguns conjuntos de dados de referência, o problema dos vendedores ambulantes e a bioinformática, um novo campo atraindo esforços intensivos. Vários temas bem relacionados, medidas de proximidade e validação de cluster também são discutidos.

(SRIVASTAVA, A.; SOTO, A. J.; MILIOS, E., 2013) faz um estudoda utilização de uma projeção em modo único do grafo bipartido de co-ocorrência de palavras-chaves e a posterior aplicação de um algoritmo de otimização de modularidade para agrupar os documentos. Os autores mostram que os algoritmos baseados em projeções de modo único funcionam significativamente melhor do que as abordagens de cluster tradicionais.

(NIU, D.; DY, J.; JORDAN, M., (2014)relatam que dados complexos são agrupados e interpretados de várias maneiras diferentes e que a maioria dos algoritmos de agrupamento existentes podem fornecer pouca orientação aos analistas de dados que por não estarem satisfeitos com um cluster único podem desejar explorar alternativas. Apresentam uma abordagem inovadora que fornece várias soluções de cluster para o usuário para fins de análise de dados exploratórios. Desenvolvem um algoritmo baseado em um procedimento de otimização que incorpora termos de qualidade de cluster e novidade em relação ao agrupamento descoberto.

(GUEDES, G. P.; OGASAWARA, E.; *et al*., 2016) esse trabalho os autores apresentam um novo algoritmo para gerar agrupamentos múltiplos a partir de um grafo com atributos. Nesse tipo de grafo

, cada vértice está associado a uma *n*-tupla de atributos (por exemplo, em uma rede social, os usuários têm interesses, sexo, idade, etc.). A abordagem concebida adiciona arestas artificiais entre vértices semelhantes do grafo utilizando a similaridade entre os atributos, o que resulta em um grafo com atributos aumentado.

(GARIN, G. L.; EMMENDORFER, L. R., 2017), apresentam o agrupamento espectral baseado em uma etapa de aglomeração dos k-menores autovalores da matriz Laplaciana que representa o conjunto de dados no grafo. Desenvolvem um algoritmo de aglomeração com uma proposta de contornar o problema causado pelo algoritmo tradicional que utiliza o k-médias para estabelecer um agrupamento nos k-menores autovetores da matriz Laplaciana.

CAPÍTULO 3.

Esse capítulo aborda conceitos fundamentais dos algoritmos de agrupamento particional utilizados no presente trabalho.A Seção 3.1 começa apresentando o algoritmo mais conhecido na literatura o k-médias. A seguir, são apresentados os conceitos matemáticos do tradicional algoritmo de agrupamento espectral para o entendimento dos algoritmos desenvolvidos por (Ng. A. Y.;*et al.*, 2001) e (GARIN, G. L.; EMMENDORFER, L. R., 2017). Ressaltando que a medida de qualidade utilizada para esse trabalho, para analisar a performance dos algoritmos é a medida F.

3. FUNDAMENTAÇÃO TEÓRICA

3.1 ALGORITMOS DE PARTICIONAMENTO

Os algoritmos de agrupamento particional podem ser aplicados em grandes quantidades de objetos e seus conceitos serão utilizados no decorrer desse trabalho.

O algoritmo k-médias é um dos algoritmos particionais mais utilizados e mais conhecidos (HAN *et al*., 2011). (LAURENT *et al*., 2014), afirma que é amplamente utilizado devido a sua simplicidade e competência. O k-médias é iniciado com a escolha dos centroides (centros dos grupos), que são pontos no espaço de objetos que representam uma posição média em cada grupo (SAJJA P. S.; AKERKAR, R. 2012). O Algoritmo 1 descreve os passos do k-médias.

Algoritmo 1: Algoritmo k-médias (D, *k*)

Entrada: Um conjunto de dados D contendo *n* objetos.

k = número de grupos.

Saída: Um conjunto de k grupos.

1: Escolher aleatoriamente *k* objetos de D como os centroides iniciais de cada grupo.

2: Calcular a distância entre cada objeto e os centroides, adicionando o objeto ao grupo que possuir menor distância.

3: Atualizar a média dos grupos, ou seja, calcular a média dos valores dos objetos para cada grupo (centroides).

4: Repetir os passos 2 e 3 até que não haja mais mudança.

O objetivo do k-médias é buscar minimizar, de forma iterativa, a distância entre os *n* objetos e os *k* centros. O Passo 1 do algoritmo escolhe aleatoriamente *k* objetos para serem os centroides dos grupos. No Passo 2, cada um dos objetos restantes é associado ao grupo ao qual mais se assemelha, baseado na distância euclidiana entre o objeto e o centroide dos grupos. A distância euclidiana entre dois pontos é calculada conforme a Eq. 5. Nessa equação, x_ie $y_i$são os pontos no espaço Euclidiano e *s* é o número de dimensões.

$$d = \sqrt{\sum_{i=1}^{s}(x_i - y_i)^2} \tag{9}$$

No Passo 3, o algoritmo atualiza os valores dos centroides de cada grupo utilizando a média aritmética dos objetos assinalados a cada grupo. O k-médias determina um número *k* de partições tentando minimizar uma função objetivo. A função objetivo mais comumente utilizada é a soma dos erros quadráticos (Sum of SquaredError – em inglês)(*SSE*), conforme apresentado na Eq. 6. Nessa equação, *k* é o número de grupos, *x* é um ponto no grupo g_i e $cent_i$ é o centróide do grupo g_i.

$$SSE = \sum_{i=1}^{k}\sum_{x \in g_i} dist^2\,(cent_i, x) \tag{10}$$

O algoritmo tradicional depende de um parâmetro (k =número de clusters) definido pelo usuário, o que poderá ser um problema, tendo em vista que normalmente não se sabe quantos clusters existem a priori. A Figura 7, ilustra os passos do Algoritmo 1 para um conjunto de *n* objetos e *k* = 3.

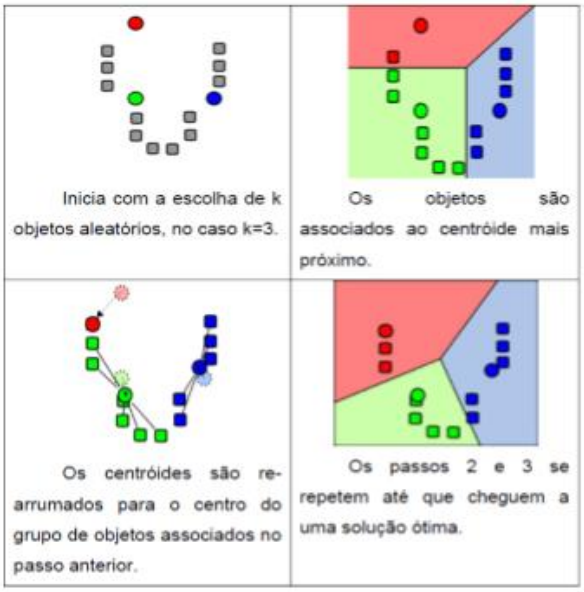


Figura 7:Exemplo k-médias
Fonte: ABREU, N.M.M.; OLIVEIRA, C.S., 2004.

O algoritmo 1 é extremamente veloz, geralmente convergindo em poucas iterações para uma configuração estável, na qual nenhum elemento será designado para um grupo cujo centro não lhe seja o mais próximo. No entanto, um eventual problema é que esta condição enfatiza a questão da homogeneidade e ignora a importante questão da boa separação dos grupos (clusters). Isto pode causar uma má separação dos conjuntos no caso de uma má inicialização dos centros, inicialização esta que é feita de forma arbitrária (aleatória) no início da execução. Na Figura 8, pode-se ver o efeito de uma má inicialização na execução de um algoritmo de k-médias. Outro ponto que pode afetar a qualidade dos resultados é a escolha do número de conjuntos feita pelo usuário. Um número muito pequeno de conjuntos pode causar a junção de dois clusters, enquanto que um número muito grande pode fazer com que um cluster natural seja quebrado artificialmente em dois.

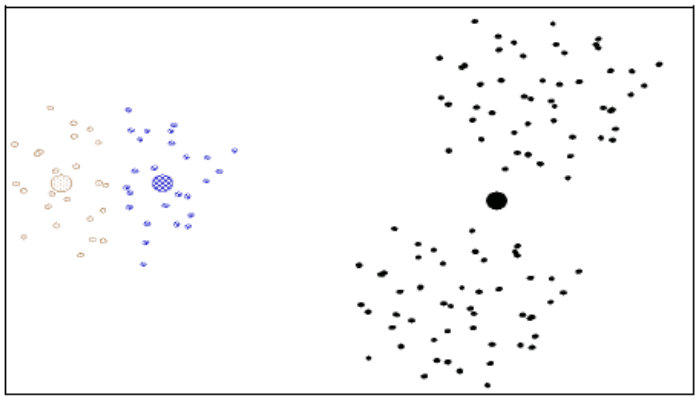


Figura 8:Exemplo do efeito de uma má inicialização do algoritmo k-médias. Existem três clusters naturais neste conjunto, dois dos quais foram atribuídos ao grupo da esquerda. O problema é que o terceiro cluster (direita) é bem separado dos outros do dos outros dois clusters, não conseguindo agrupar, ficando afastados do centroide.
Fonte: LINDEN, R.2005.

A maior desvantagem deste algoritmo é a necessidade de selecionar um número de grupos *k* previamente, o que exige que se saiba a priori quantos grupos tem o banco de dados ou então executar o algoritmo diversas vezes variando o valor de *k* até encontrar uma partição ideal. Desta maneira entende-se que o valor de *k* é extremamente importante e depende diretamente do algoritmo (MICHAEL, J. A. B; GORDON L.1997).

Os algoritmos de agrupamento baseado em grafos representam os dados e sua proximidade através de um grafo *G* (*V*, *E*), onde *V*= $\{u_1, u_2, u_3, \ldots . u_q\}$representam os vértices e *E* representa o conjunto de arestas. Cada vértice representa um elemento do conjunto de dados e a existência de uma aresta unindo dois vértices é feita com base na proximidade entre os dois objetos. A maneira mais simples de estabelecer as ligações entre os vértices é conectar cada vértice aos (n-1) vértices restantes, onde o peso da aresta indica a similaridade entre os dois objetos que a mesma conecta.

Em um grafo simples, apenas uma aresta pode unir dois vértices $u_1 e u_2$. Além disso, um mesmo vértice não pode ser origem e destino ao mesmo tempo. Caso ao menos uma dessas duas condições não seja respeitada, esse grafo é denominado um multigrafo.

Em um grafo não-dirigido, as arestas não possuem orientação, de forma que *E* é um conjunto de subconjuntos de dois elementos de *V*. Caso o conjunto *E* seja formado por pares ordenados de arestas (u_1,u_2) apresenta-se um grafo dirigido. É importante ressaltar que o foco desse trabalho ocorre em grafos simples não-dirigidos. As Figuras

9(a) e 9(b)apresentam um grafo simples não-dirigido e um grafo simples dirigido respectivamente.

(a) Grafo simples não-dirigido sem pesos nas arestas.

(b) Grafo simples dirigido sem pesos nas arestas.

Figura 9: Exemplo de grafo não-dirigido e grafo dirigido.

Os grafos podem ser representados através de suas matrizes de adjacência para análises matemáticas. Considerando um grafo simples não-dirigido, a matriz de adjacência *A* e seus elementos A_{ij} são representados da seguinte forma:

$$A = \begin{cases} 1 & se\ \ existe\ uma\ aresta\ entre\ os\ vértices\ u_i\ e\ u_j \\ 0 & caso\ contrário \end{cases}$$

$$A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

Figura 10:Matriz de adjacência para o grafo representado na Figura 9a.

Além da matriz de adjacência, pode-se calcular a matriz de grau, que representa um aspecto importante na análise de um grafo. É determinada pelo número de arestas incidentes a cada vértice. Essa matriz pode ser calculada a partir da matriz de adjacência conforme a Eq. 7, em que d_i denota o grau do vértice u_i.

$$d_i = \sum_{i=1}^{n} A_{ij} \tag{11}$$

A matriz de grau D pode ser calculada utilizando a Eq. 8.e possui a seguinte definição:

$$D_{ij} = \begin{cases} d_i & se\ \ i = j \\ 0 & caso\ contrário \end{cases} \tag{12}$$

A partir da definição acima, a matriz de grau referente à matriz de adjacência ilustrada na Figura 10 é apresentada na Figura 11. Pode-se notar que a matriz *D* é uma matriz diagonal.

$$A = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Figura 11:Matriz de grau para o grafo representado na Figura 10

As matrizes de adjacência e diagonal são importantes na obtenção de grupos. A tarefa de agrupamento em grafos consiste em agrupar vértices do grafo em grupos, considerando a estrutura das arestas, de forma que exista muitas arestas dentro de cada grupo e relativamente poucas entre os grupos. Existem diversos algoritmos de agrupamento em grafos, dentre eles podemos destacar o algoritmo de agrupamento espectral. A teoria espectral tem como um de seus principais objetivos deduzir as principais propriedades e a estrutura de um grafo a partir do seu espectro (CHUNG, 1997). Devido à sua eficiência e desempenho, o algoritmo de agrupamento espectral se tornou um dos mais populares métodos de agrupamento (CHUANG, 2012). A principal componente dos algoritmos de agrupamento espectral é a matriz Laplaciana (L), definida conforme a Eq. 9. Nessa equação, D representa a matriz de grau e A é a matriz de adjacência de um grafo.

$$L = D - A \tag{13}$$

A Figura 12 apresenta a matriz Laplaciana para o grafo ilustrado na Figura9(a).

$$A = \begin{bmatrix} 2 & -1 & -1 \\ -1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

Figura 12:Matriz Laplaciana para o grafo representado na Figura 9a.

O agrupamento espectral tem o objetivo de particionar as linhas de uma matriz de acordo com as componentes dos autovetores associados à matriz. Assim, dada a matriz Laplaciana L associada a um grafo G, cada vértice é representado em um espaço vetorial usando alguns autovetores associados a essa matriz. Esses autovetores são apresentados em uma nova matriz, cujas linhas representam os pontos no Espaço Vetorial e as colunas representam os autovetores. Segundo,(VON LUXBURG, 2007), o Algoritmo 2 descreve os passos necessários para o agrupamento tradicional espectral.

Algoritmo 2: Algoritmo tradicional de agrupamento espectral (G, k)

Entrada: G (V, E) = grafo com um conjunto de vértices V e um conjunto de arestas E.
k= número de grupos.
Saída: $\{V_i\}_{k=1}^{n}$, uma partição de V.
1: Computar a matriz de adjacência A.
2: Computar a matriz de grau D.
3: Computar a matriz Laplaciana $L = D - A$.
4: Encontrar os $k$menores autovetores de L e formar a matriz $Q = [q_1, ...q_k]$.
5: Considerando cada linha de Q como um ponto no Espaço Vetorial, agrupá-los em $k$grupos usando o algoritmo k-médias.
6: Dado um vértice $x_i \in G$, assinalar x_i ao grupo j se, e apenas se, a linha i de Q é assinalada ao grupo j.

O peso de cada aresta representa a similaridade entre os elementos de acordo com alguma medida (WILLIAM, M, R. 1971) 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. A partir dessa matriz, aplica-se um algoritmo de agrupamento padrão, o k-médias, sobre essas novas coordenadas (Luz, D. 2003). De acordo com (XU, R.; WUNSCH, D., I. 2005), 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 poderá ser feito 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.

O Algoritmo 2 modificado por (Ng. A. Y.; *et al.*, 2001),no passo 1 ao invés de computar a matriz de adjacência A, computa a matriz de similaridade

$$W_{ij} = e^{\left(\frac{-1}{2\sigma^2} d^2(x_i, x_j)\right)} \tag{14}$$

onde $e^{\left(\frac{-1}{2\sigma^2} d^2(x_i, x_j)\right)}$ é a função gaussiana e sua utilização é para normalização dos dados. No passo 3, computa a matriz Laplaciana normalizada

$$L = D^{\frac{-1}{2}}(D - W)D^{\frac{-1}{2}} \quad (15)$$

Os passos do Algoritmo 2, por(Ng. A. Y.; *et al.*, 2001), basicamente desempenha a tarefa de obter a matriz Laplaciana normalizada do grafo de similaridade e calcular os seus autovetores para agrupá-los via k-médias. O parâmetro σ^2, controla a rapidez com que os elementos da matriz de similaridade A_{ij} decresce quando leva em consideração a distância entre x_ie x_j.

Recentemente,(GARIN, G. L.; EMMENDORFER, L. R., 2017) desenvolveram um algoritmo baseado no Algoritmo espectral via k-médias por (A. Ng, M. J.; Y.Weiss, 2001). Nessa nova versão,motivados pela ineficiência do método em conjuntos de dados de formato não esférico, o algoritmo de agrupamento espectral não tem a etapa do mapeamento dos autovetores utilizando o algoritmo k-médias. Apesar da convergência do k-médias ocorrer em poucas iterações, uma má inicialização dos centroides torna-se um problema, afetando o resultado do agrupamento. Isto ocorre porque os centroides são inicializados aleatoriamente. Se existem *k* grupos reais, então a probabilidade de selecionar um centroide para cada grupo é relativamente pequena (TAN, P-N; *et al.*2009). Além disso, esta metodologia é sensível ao parâmetro de vizinhança estabelecido na construção do grafo de representação dos dados. Desse modo, uma abordagem por aglomeração na etapa de mapeamento dos autovetores pode se tornar um método alternativo para contornar os problemas existentes com a abordagem do k-médias. O Algoritmo 3 descreve os passos do Algoritmo de Agrupamento Espectral Aglomerativo

---

Algoritmo 3: Algoritmo de agrupamento espectral aglomerativo (*G*, *k*)

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

Saída: Grupos rotulados $A_1, \ldots, A_k$

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

2: Computar a matriz de grau *D*.

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

4: Encontrar os k autovetores de L (escolhidos para serem ortogonais entre si no caso de autovalores repetidos), e formar a matriz U colocando os autovetores em colunas: $U = [u_1 \vdots \cdots \vdots u_2] \in \mathbb{R}^{n\,x\,k}$

5: Computar a matriz Y a partir de U normalizando cada linha de U para ter valores unitários.

6: Computar 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, ou seja, $d(Y_a, Y_b) = \frac{1}{|Y_a||Y_b|}\sum_{y_i \epsilon Y_a}\sum_{y_j \epsilon Y_b} dist(y_i, y_j)$

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

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

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

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

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

$$\boldsymbol{if}\ rot(a) \neq 0\ and\ dist(a,b) \cdot numviz(a) \leq dviz(a)\boldsymbol{then}$$

$$\begin{cases} dviz(b) += dist(a,b); \\ numviz(b) += 1; \\ rot(b) = rot(a); \end{cases}$$

$$\boldsymbol{if}\ numviz(b) \geq viz\ \boldsymbol{then}$$

$$\begin{cases} rot(b) = 0; \\ numviz(b) = 1; \\ dviz(b) = d(Y_a, Y_b); \end{cases}$$

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

A busca pelo número de grupos que acontece nos autovetores é feita de maneira aglomerativa dentro desse novo conjunto. A informação que os autovetores obtidos, no passo 4 do algoritmo, 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.

No passo 6, uma matriz de distâncias é gerada de acordo com as coordenadas dos autovetores e no passo 7 é determinada a média das distâncias. No passo 8, 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. No passo 9, são selecionadas duas posições aleatórias do conjunto dos autovetores, onde cada posição o 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 eb 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 uma posição é utilizada uma comparação entre o número de vizinhos desta posição e o número de vizinhos do conjunto de dados atual (autovetores). Se a quantia de vizinhos da posição b extrapolar o número de vizinhos do conjunto de dados, então o processo de aglomeração o naquela posição termina. A parada total do método ocorre no passo 10, quando todas as posições dos vetores forem submetidos ao teste do passo 9.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.Ao final da execução do Algoritmo Aglomerativo, a quantidade de rótulos é igual ao número de grupos do conjunto. A Figura 13 mostra um grafo com quatro grupos distintos.

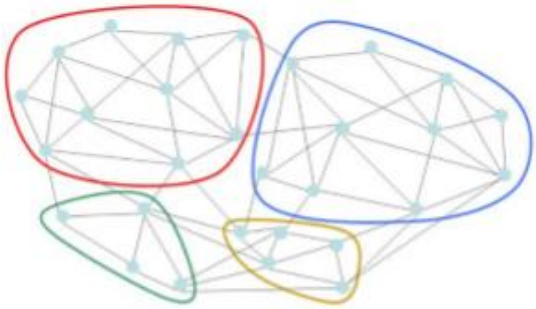


Figura 13: Grafo Particionado em 4 Grupos
Fonte: **GARIN, G. L. 2017**

Ressalta-se que a distância entre grupos também pode ser medida de acordo com outras métricas, como por exemplo, as métricas apresentadas na seguinte tabela:

Tabela 1: Distância entre Grupos Usados em Diferentes Algoritmos Aglomerativos

Método	Distância entre grupos
Single-link	$\boldsymbol{d(C_a, C_b) = min_{ij} d(x_i \epsilon C_a, x_j \epsilon C_b)}$

Complete-link	$\boldsymbol{d(C_a, C_b) = max_{ij} d(x_i \epsilon C_a, x_j \epsilon C_b)}$
Average-link	$\boldsymbol{d(C_a, C_b) = \frac{1}{\lvert C_a \rvert \lvert C_b \rvert} \sum_{x_i \epsilon C_a} \sum_{x_j \epsilon C_b} d(x_i, x_j)}$
Centroid-link	$\boldsymbol{d(C_a, C_b) = d(C_a, C_b)}$

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 baseia-se na informação global do conjunto, assim, não havendo a preocupação o 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 $Centroid - link$ promissor quando os grupos são globulares e de mesma densidade, ou seja, tal critério funciona bem apenas em um conjunto limitado de problemas.

CAPÍTULO 4.

Este capítulo é destinado aos testes computacionais realizados, resultados e discussão. Também é apresentada as considerações finais sobre o trabalho realizado.

4. TESTES COMPUTACIONAIS

Três testes computacionais foram realizados sobre um total de 10 bancos de dados, disponíveis no site (http://stackoverflow.com/questions/16146599/create-artificial-data-inmatlab) de diferentes formatos geométricos e bidimensionais, utilizando o algoritmo desenvolvido Aglomerativo, k-médias e Espectral via k-médias. Houve a preocupação, em trabalhar com conjuntos de dados de diferentes formatos geométricos para analisar os resultados de cada algoritmo testado. Cada algoritmo foi rodado 10 vezes para cada banco de dados, no intervalo [0,1], onde o valor 1 é considerado resultado ótimo para medida-F.

4.1. TESTE 1 (k-médias)

Nesse teste, o algoritmo utilizado foi o k-médias. A Tabela 2, mostra a descrição dos bancos de dados e o valor da medida-F. O algoritmo foi rodado 10 vezes, sendo $\boldsymbol{F_n}$ a notação utilizada para cada rodada do algoritmo. Cabe salientar que, no teste do algoritmo k-médias, não existe a possibilidade de alteração do número de vizinhos.

Tabela 2: Algoritmo k-médias (rodado dez vezes para cada banco de dados)

Nome do Banco de Dados	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	F_{10}
Atom	1	1	1	1	1	1	1	1	1	1
Chainlink	1	1	1	1	1	1	1	1	1	1
ClusterinCluster	0,5778	0,5333	0,4444	0,4667	0,4222	0,5333	0,6000	0,5333	0,5778	0,5556
EngyTime	1	1	1	1	1	1	1	1	1	1
Flame	0,8276	0,0192	0,8060	0,0192	0,8276	0,0284	0,0284	0,8259	0,8060	0,0563
Halfkernel	0,4660	0,4660	0,4660	0,0540	0,4740	0,5260	0,5260	0,5260	0,5260	0,5260
Jain	0,3540	0,3540	0,8288	0,3540	0,3540	0,3540	0,8288	0,3540	0,8288	0,8288
Twodiamonds	1	1	1	1	1	1	1	1	1	1
Twospirals	0,3770	0,6230	0,6230	0,6230	0,6230	0,6230	0,6230	0,6230	0,3770	0,3770
WingNuts	1	1	1	1	1	1	1	1	1	1

Através da Tabela 2, é possível observar que o conjunto de dados Atom, Chainlink, EngyTime, Twodiamonds e WingNuts, obtiveram resultado ótimo. O conjunto de dados Jain e Flame, com medida-F de 0,8288 e 0,8276, respectivamente. O conjunto de dados Twospirals e ClusterinCluster apresentaram medida-F de 0,6230 e 0,6000. O conjunto de dados Halfkernel, com medida-F de 0,5260.

4.2. TESTE 2. (Espectral via k-médias)

Nesse teste, considera-se a possibilidade de alternar o número de vizinhos de um elemento, ou seja, o número de vizinhos é informado a cada execução do algoritmo. 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 fato, 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. A Tabela 3, considera o número de vizinhos entre 10 e 100, e 30 apresenta os seguintes resultados:

Tabela 3: Espectral via k-médias (alterando o número de vizinhos)

Nome doBanco de Dados	F_1 (v=10)	F_2 (v=20)	F_3 (v=30)	F_4 (v=40)	F_5 (v=50)	F_6 (v=60)	F_7 (v=70)	F_8 (v=80)	F_9 (v=90)	F_{10} (v=100)
Atom	1	1	1	1	1	1	1	1	1	1
Chainlink	1	1	1	1	1	1	1	1	1	1
ClusterinCluster	0,6723	0,6723	0,6723	0,0020	0,0020	0,6723	0,6000	0,0020	0,5778	0,6723
Engytime	1	1	1	1	1	1	1	1	1	1
Flame	0,8276	0,7834	0,0199	0,0193	0,0365	0,0291	0,8000	0,7873	0,7699	0,0112
Halfkernel	1	1	0,6667	0,5185	0,3593	0,4289	0,3779	0,4215	0,4165	0,3871
Jain	0,3311	0,7310	0,6504	0,5817	0,6667	0,5286	0,5139	0,3540	0,5020	0,4990
Twodiamonds	1	1	1	1	1	1	1	1	1	1
Twospirals	1	0,6667	0,5730	0,1933	0,4171	0,4756	0,3444	0,3498	0,4592	0,2979
WingNuts	1	1	1	1	1	1	1	1	1	1

Através da Tabela 3, pode-se observar que os bancos de dados Atom, Chainlink, EngyTime, Twodiamonds e WingNuts, obtiveram resultado ótimo. O banco de dados Halfkernel, com coeficiente de variabilidade 44,75% tem um resultado ótimo para 10 e

20 vizinhos. O banco de dados Twospiral, com resultado ótimo somente para 10 vizinhos. Os outros banco de dados não apresentaram bons resultados.

No entanto, analisando o gráfico representado na Figura 14, percebe-se que os resultados obtidos para 20 vizinhos (cor laranja), foi quando o conjunto de dados obteve os melhores resultados.

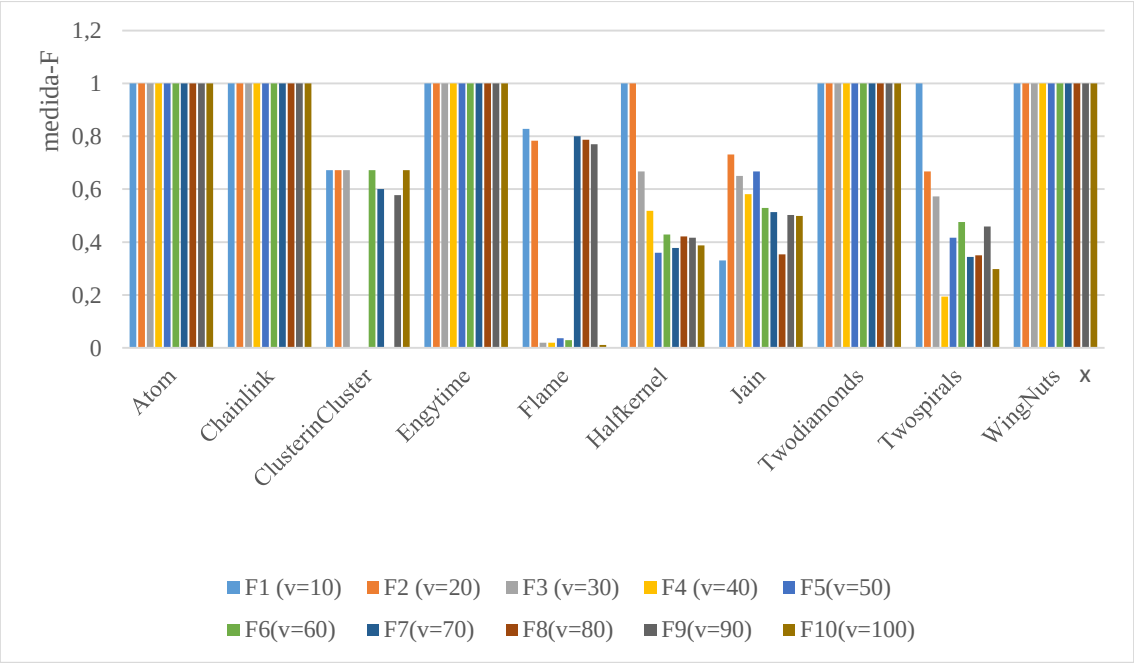


Figura 13: Gráfico Teste Espectral via K-médias.

A fim de observar,o comportamento do conjunto de banco de dados analisados anteriormente, para v = 20 vizinhos o algoritmo foi novamente rodado.A Tabela 4, mostra os resultados obtidos.

Tabela 4: Espectral via-k-médias (20 vizinhos)

Nome do Banco de Dados	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	F_{10}
Atom	1	1	1	1	1	1	1	1	1	1
Chainlink	1	1	1	1	1	1	1	1	1	1
ClusterinCluster	0,9980	0,9980	0,9980	0,9980	0,9980	0,6723	0,6723	0,9980	0,6723	0,9980
Engytime	1	1	1	1	1	1	1	1	1	1
Flame	0,0203	0,7834	0,0203	0,0203	0,7834	0,7834	0,0203	0,0203	0,0203	0,0203
Halfkernel	1	1	0,6667	1	0,6667	1	0,5260	0,5260	0,5260	0,5260
Jain	0,4776	0,7310	0,4776	0,4776	0,4776	0,4776	0,4776	0,7310	0,7310	0,7310
Twodiamonds	1	1	1	1	1	1	1	1	1	1
Twospirals	0,6667	0,6667	0,6667	0,6667	0,6667	0,6667	0,6667	0,6667	0,6667	0,6667
WingNuts	1	1	1	1	1	1	1	1	1	1

Através da Tabela 4, é possível observar que os bancos de dados Atom, Chainlink, EngyTime, Twodiamonds e WingNuts, obtiveram resultado ótimo. Os bancos de dados Halfkernel, ClusterinCluster apresentaram uma pequena dispersão, com coeficiente de variabilidade 30,53, 16,83% respectivamente, apresentando melhores resultados quando comparados ao teste anterior. O banco de dados Twospirals não apresentou variação em seus resultados. Os bancos de dados Flame e Jain conseguem obter medida-F máxima igual a 0,7834 e 0,7310 respectivamente.

4.3. TESTE 3 (Aglomerado)

O algoritmo aglomerado, diferentemente dos algoritmos apresentados, não altera o valor da medida-F. Mesmo quando rodado várias vezes, a medida-F é mantida.

Tabela 5: Aglomerado

Atom	1,00000
Chainlink	1,00000
ClusterinCluster	0,99803
Engytime	1,00000
Flame	0,50000
Halfkernel	0,66667
Jain	1,00000
Twodiamonds	1,00000
Twospirals	1,00000
WingNuts	1,00000

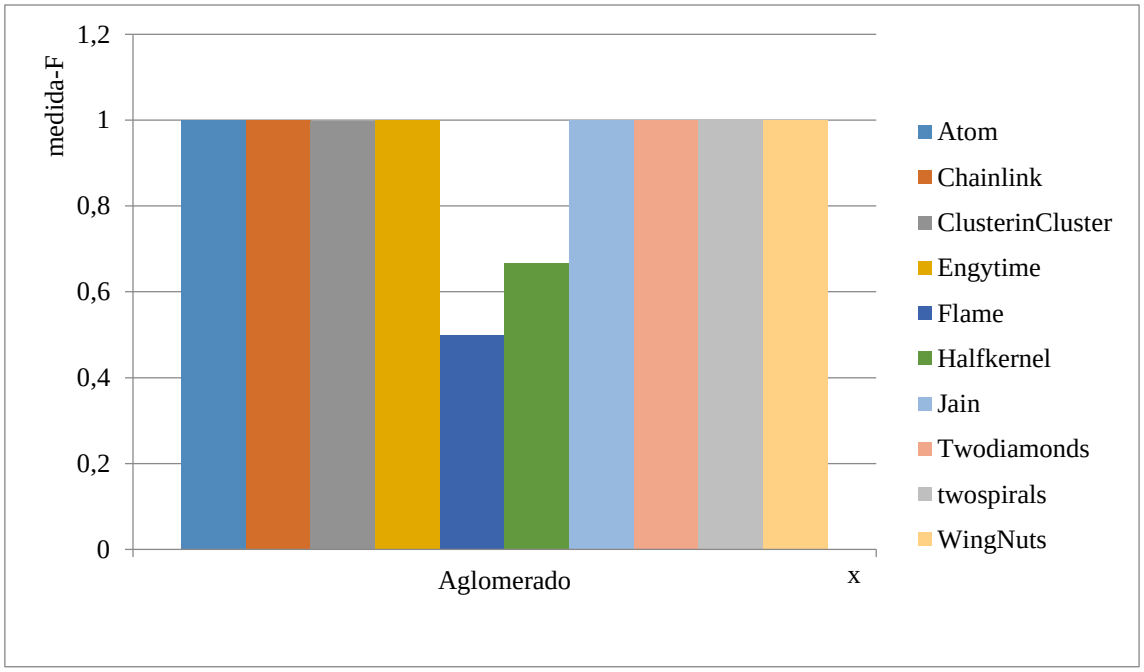


Figura 14: Gráfico Teste Aglomerado

Assim, o algoritmo aglomerado obteve uma média aritmética 0,91647 com uma variabilidade de 25,31%, obtendo um resultado melhor que os demais algoritmos testados.

Com relação ao banco de dados Jain, é possível mostrar que os resultados do algoritmo Aglomerativo (medida-F = 1)é superior aos resultados dos algoritmos k-médias (no pior caso, com medida-F = 0,3540) e espectral via k-médias (medida-F = 0,6504). Através da Tabela 6, os resultados podem ser verificados.

Tabela 6: Comparação Através de Figuras

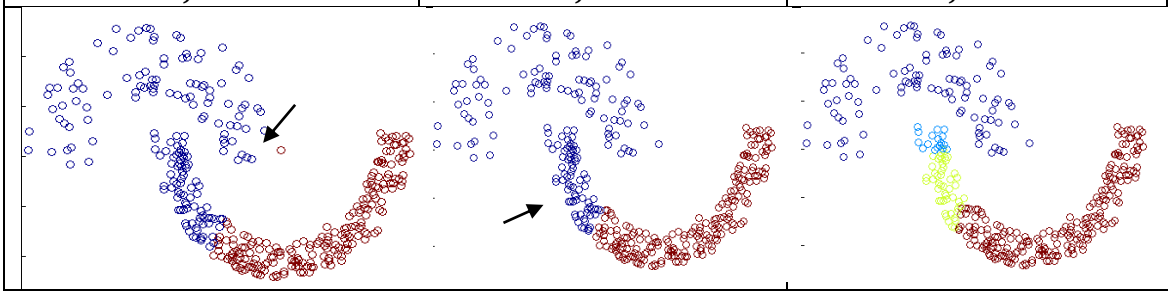
k-médias	Espectral via-k-médias	Aglomerado
0,3540	0,6504	1,0000
		

Figura 15: Comparação através de figuras (Banco de dados Jain).

Pode ser observado através da Figura 16, que na primeira imagem, existem dois agrupamentos: um agrupamento identificado somente pela cor azul e o outro agrupamento identificado pelas cores azul e vermelho. Pode-se dizer que o algoritmo k-médias acertou 35,40% do agrupamento, pois existe um dado completamente isolado

(em vermelho) apontado através da flecha. Além disso, no segundo agrupamento aparecem grupos misturados.

Na segunda imagem, o algoritmo espectral via k-médias acertou 65,04 % posicionando o dado, que anteriormente estava posicionado no grupo errado, corretamente. No entanto, o segundo algoritmo (espectral via-k-médias), continua errando o segundo agrupamento, misturando os dados (cor azul e vermelho).

Na terceira imagem, o algoritmo aglomerado consegue separar os dados corretamente em quatro grupos, obtendo assim um resultado considerado ótimo quando utilizada a medida-F.

Pode-se concluir que, quando o conjunto de dados possui determinada curvatura, como é o caso do conjunto Jain, a utilização dos centroides como parâmetro de agrupamento não funciona. Assim, o algoritmo k-médias não consegue um agrupamento coerente de acordo com a estrutura dos dados. Portanto, a distribuição dos dados no plano não acompanha uma distribuição Gaussiana, então o uso do k-médias neste caso, não é recomendado.

Nos grupos obtidos pelo algoritmo espectral via k-médias o resultado foi melhor que o k-médias. No entanto, a informação da matriz Laplaciana não foi suficiente para estabelecer um agrupamento coerente com a geometria do conjunto. Pode-se notar que os resultados do último algoritmo de agrupamento foram satisfatórios, respeitando a estrutura organizacional dos dados. A explicação para este fato é decorrente de que o método utilizado não é baseado em um conjunto de centroides, mas em uma relação entre os elementos de acordo com a sua média dentro do conjunto de autovetores.

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-F.

4.4. CONSIDERAÇÕES FINAIS

Nesse trabalho foram apresentados experimentos com dados disponíveis na literatura para aplicação dos algoritmos espectrais e k-médias. Os resultados obtidos através da medida-F, mostraram que o algoritmo aglomerativo é promissor para futuras aplicações.

Uma das vantagens do método por aglomeração é a menor sensibilidade ao número de k-vizinhos escolhidos para montar o grafo de similaridade. O método de aglomeração varia os resultados do agrupamento, conforme o aumento do número de vizinhos.

Para trabalhos futuros, uma investigação maior é recomendada com relação ao conjunto de dados Flame e Halfkernel. Recomenda-se também relacionar os resultados obtidos pelo agrupamento espectral aglomerativo, através da medida-F, com um parâmetro estatístico.

5. REFERÊNCIAS BIBLIOGRÁFICAS

ANDA, C. **Data Mining Techniques in Supporting Decision Making**.1999.

ABREU, N. **Introdução a Teoria Espectral de Grafos com Aplicações**. São Carlos, SP: SBMAC, 4, 15. 2007.

ABREU, N.M.M.; OLIVEIRA, C.S. **Álgebra Linear em Teoria dos Grafos**, mini-curso, Primeira Semana da Matemática de São Mateus, ERMAC/SBMAC, UFES. 2004.

BARBAKH, W. A.; WU, Y.;FYFE, C. **Review of clustering algorithms.**Studies in ComputationalIntelligence, 249:7–28. 2009.

BOAVENTURA NETTO, P. O. **Grafos: teoria, modelos e algoritmos**. São Paulo: EdgardBlücher. 2003.

BOTHOREL, C., CRUZ, J. D., MAGNANI, M., et al..**Clustering attributed graphs: models, measures and methods**, CoRR, v. abs/1501.01676. Disponívelem:<https://arxiv.org/abs/1501.01676>. 2015.

BURK, I. **Spectral clustering**. In Bachelor Thesis-University of Stuttgart. 2012.

CASSIANO, K. M. Análise de Séries Temporais Usando Espectral Singular (SSA) e Clusterização de suas Componentes Baseadas em Densidades. Tese de Doutorado. Pontíficia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica, 2014.

CHANG, H.; YEUNG. D. Y. H. Robust path-based spectral clustering. In Pattern Recognition, pages 191–203. 2008.

CHUNG, F.**Spectral Graph Theory**. American Mathematical Society, National Science Foundation.1997.

CONCEITOS BÁSICOS DA TEORIA DE GRAFOS. Disponível em: http://www.inf.ufsc.br/grafos/definicoes/definicao.html. Acesso em: 21/08/2015.

CVETKOVIC, D.; ROWLINSON, P.; SIMIC, S.**Eigenspaces of Graphs in: Encyclopedia of Mathematics and its Applications**, Cambridge, vol. 66. 1997.

CHUANG, Y.-Y.**Affinity Aggregation for Spectral Clustering**. In: Proceedings. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), CVPR '12, pp. 773–780, Washington, DC, USA. IEEE Computer Society. ISBN: 978-1-4673-1226-4. Disponível em:http://dl.acm.org/citation.cfm?id=2354409.2355074. 2012.

CHUNG, F. Spectral Graph Theory, ser. **Regional Conference Series in Mathematics**. Conference Board of the Mathematical Sciences, no. No 92. [Online]. Available: http://books.google.com.br/books?id=YUc38_MCuhAC. 1997.

DAVIES, D. L., BOULDINB, D. W., **A Cluster separation measure**, IEEE Transactions on Pattern Analysis and Machine Intelligence, v. PAMI-1, pp. 224-227, 1979.

FRITSCHER, E. **Propriedades Espectrais de um Grafo.** 126f. Dissertação de Mestrado em Modelagem Aplicada – Universidade Federal do Rio Grande do Sul. 2011.

GAN, G., MA, C., WU, J.**Data clustering - theory, algorithms, and applications**. SIAM. 2007.

GARIN, G. L.; EMMENDORFER, L. R. **Agrupamento Espectral Aglomerativo: Uma Prosposta de Algoritmo**. Dissertação de Mestrado em Engenharia de Computação – Universidade Federal do Rio Grande do Sul. 2017.

GORDON, A. D., Classification, Chapman and Hall Ed.1981.

GUEDES, G. P.; OGASAWARA, E.; BEZERRA, E. ; XEXEO, G. **Discovering top-k non-redundant clusterings in attributed graphs**. NEUROCOMPUTING, v. 210, p. 45-54. 2016.

HAN, J., KAMBER, M., PEI, J.**Data Mining: Concepts and Techniques**. 3rd ed. San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. ISBN: 0123814790, 9780123814791. 2011.

HARTIGAN, J. A.; WONG, M. A. **A K-means clustering algorithm: Applied Statistics**, 28:100–108. 1979.

INGO, B. **Spectral clustering.** In Bachelor Thesis-University of Stuttgart. 2012.

JAIN, A.K.; DUBES, R.C. **Algorithms for Clustering Data,** Prentice Hall, 1988.

KURSUN, O. (2010). **Spectral Clustering with Reverse Soft K-Nearest Neighbor Density Estimation**. In: *International Joint Conference on NeuralNetworks, IJCNN 2010, Barcelona, Spain, 18-23 July*, pp. 1–8. doi: 10.1109/IJCNN.2010.5596620. Disponívelem: <http://dx.doi.org/10.1109/IJCNN.2010.5596620>. 2010.

KUNCHEVA, L. I. **Combining pattern classifiers: methods and algorithms.** John Wiley & Sons, 2004.

LAURENT, A., STRAUSS, O., BOUCHON-MEUNIER, B., et al., (2014), Information Processing and Management of Uncertainty: 15th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU, Montpellier, France, July 15-19, 2014. Proceedings. N. pt. 1, Communications in Computer and Information Science. Springer International Publishing. ISBN: 9783319087955. Disponível em:https://books.google.com.br/books?id=IWIlBAAAQBAJ .2014.

LINDEN, R.**Um Algoritmo Híbrido Para Extração De Conhecimento Em Bioinformática**. 213 f. Tese(Doutorado em Ciências em Engenharia Elétrica) – Universidade Federal do Rio de Janeiro. 2005.

LUZ, D.**Implementação o de uma Ferramenta de Data Mining para o Auxílio à Tomada de Decisão - Caso de uma Cadeia de Suprimentos.** Trabalho de Conclusão de Curso - Engenharia de Controle e Automação Industrial, Universidade Federal de Santa Catarina - UFSC, Santa Catarina, SC, Brasil.2003.

MANNING, C.D.; RAGHAVAN, P.; SHUTZE, H.; **Introduction to Information Retrieval**, Cambridge University Press, 2008.

Machine Learning Repository. Disponívelem: https://archive.ics.uci.edu/ml/datasets.html?format=&task=clu&att=&area=&numAtt=&numIns=&type=&sort=nameUp&view=table. Acessoem 04/06/2015.

MARCO, S. **The principal components analysis of a graph, and its relationships to spectral clustering.** In European Conference on Machine Learning, pages 371–383. 2004.

MICHAEL, J. A. B; GORDON L. **Data Mining Techiques for Marketing, Sales, and Customer Support**; John Wiley & Sons, Inc., 1997.

MICHAEL, I, J.; ANDREW, Y, N.; YAIR, W. **On spectral clustering: Analysis and an algorithm**. In Advances in Neural Information Processing Systems 14. 2001.

Ng, A.Y.; JORDAN,M.I.; WEISS,Y. **On spectral clustering: Analysis and an algorithm**. In Advances in Neural Information Processing Systems 14. 2001.

NIU, D., DY, J., JORDAN, M. **Iterative Discovery of Multiple AlternativeClustering Views**, *Pattern Analysis and Machine Intelligence,IEEE Transactions on*, v. 36, n. 7 (July), pp. 1340–1353. ISSN: 0162-8828.doi: 10.1109/TPAMI.2013.180. 2014.

NUNES, B. P. **Classificação Automática de Dados Semi-Estruturados**. PUC-RIO Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, RJ, Brasil.2009.

Octave is distributed under the terms of the **GNU General Public License**.Disponível em: https://www.gnu.org/software/octave/. Acesso 09/06/15.

PEREIRA, I. A. **Uma Abordagem Transfer-learning para agrupamento de dados**. 76 f. Dissertação (Mestrado em Computação) – Universidade Federal do Rio Grande do Sul. 2013.

RAJARAMAN, A., ULLMAN, J. D.**Mining of Massive Datasets**. New York, NY, USA, Cambridge University Press. ISBN: 1107015359, 9781107015357. 2011.

ROCHA, F. C. **Métodos Espectrais para o Problema de Particionamento**. 2015. 74 f. Projeto de dissertação (Mestrado em Matemática Aplicada) – Universidade Federal do Rio Grande do Sul. 2015.

ROUSSEEUW P. J. **Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.** J. Comp. App. Math., Vol. 20, pp. 53-65, 1987.

RUFFINO, L.P.**Algoritmo de Aprendizado Supervisionado - Baseado em Máquinas de Vetores de Suportes: Uma Contribuição para o Reconhecimento de Dados Desbalanceados.** Tese de Doutorado, Pós Graduação em Engenharia Elétrica, Universidade de Uberlândia, 2011.

OVERFLOW S. http://stackoverflow.com/questions/16146599/create-artificial-data-inmatlab. 2017.

SAJJA, P. S., AKERKAR, R.**Intelligent Technologies for Web Applications**. Chapman & Hall/CRC. ISBN: 1439871620, 9781439871621.2012.

SCHAEFFER, S. E.,**Survey: Graph Clustering**.*Comput. Sci. Rev.*, v. 1, n. 1 (ago.), pp. 27–64. ISSN: 1574-0137. doi: 10.1016/j.cosrev.2007.05.001. Disponível em: <http://dx.doi.org/10.1016/j.cosrev.2007.05.001>.2007.

SRIVASTAVA, A., SOTO, A. J., MILIOS, E.**Text Clustering Using One-mode Projection of Document-word Bipartite Graphs**. In: *Proceedings of the 28th Annual ACM Symposium on AppliedComputing*, SAC ’13, pp. 927–932, New York, NY, USA. ACM. ISBN:978-1-4503-1656-9. doi: 10.1145/2480362.2480539. Disponívelem:<http://doi.acm.org/10.1145/2480362.2480539>. 2013.

STEINLEY, D., BRUSCO, M. J.,**Initializing K-means Batch Clustering: A Critical Evaluation of Several Techniques**. J. Classification, v. 24, n. 1, pp. 99–121. 2007

STEVANOVIC, D., BRANKOV, V,. CVETKOVIC, D., and SIMIC, S., **Newgraph**. Freeware. http://www.mi.sanu.ac.rs/newgraph. 2005.

TAN, P-N; STEINBACH, M.; KUMAR, V.**Introdução ao DATAMINING Mineração de Dados.**Rio de Janeiro: Editora Ciência Moderna Ltda. 2009.

VILMAR, T. NAIR, M, M, A. R. Del-Vecchio and C.T. Vinagre**. Teoria espectral de grafos uma introdução.** 2013.

VON LUXBURG, U.**A tutorial on spectral clustering, Statistics and Computing**. v. 17, n. 4, pp. 395–416. ISSN: 0960-3174. doi: 10.1007/ s11222-007-9033-z. Disponível em:<https://link.springer.com/article/10.1007%2Fs11222-007-9033-z>.2007.

XU, R.; WUNSCH, D., I. **Survey of clustering algorithms.** Neural Networks, *IEEE Transactions on*, 16(3):645–678. 2005.

WEST, D.B. **Introduction to Graph Theory**, Prentice-Hall, 1996.

WILLIAM, M, R. **Objective criteria for the evaluation of clustering methods**. Journal of the American Statistical association, 66(336):846850, 1971.

WOJCIECH, K.; KRZYSZTOF, D.; ARKADIUSZ, J.; WILLEM, W.; EYKE, H. **Optimizing the f-measure in multi-label classification: Plugin rule approach vs structured loss minimization.** In International Conference Machine Learning,pages 1130–1138. 2013.