4. Clusterização de Dados

A Clusterização de Dados ou Análise de Agrupamentos é uma técnica de mineração de dados multivariados que através de métodos numéricos e a partir somente das informações das variáveis de cada caso, tem por objetivo agrupar automaticamente por aprendizado não supervisionado os *n* casos da base de dados em *k* grupos, geralmente disjuntos denominados clusters ou agrupamentos. Na Literatura, a análise de clusters pode ser chamada também de Clusterização, Clustering, Q-analysis, Typology, Classification Analysis ou Numerical Taxonomy.

Distinta do conceito de classificação, a Clusterização é uma técnica mais "primitiva" na qual nenhuma suposição é feita a respeito dos grupos. Ao contrário da classificação, a Clusterização não conta com classes predefinidas e exemplos de treinamento de classes rotuladas, sendo assim realiza uma forma de aprendizado não supervisionado.

O primeiro registro publicado sobre um método de Clusterização foi feito em 1948, com o trabalho de SORENSEN (1948) sobre o Método Hierárquico de Ligação Completa. Desde então mais de uma centena de algoritmos distintos de clusterização já foram definidos. Qualquer método de clusterização é definido por um algoritmo específico que determina como será feita a divisão dos casos nos clusters distintos e todos os métodos propostos são fundamentados na ideia de distância ou similaridade entre as observações e definem a pertinência dos objetos a cada cluster segundo aquilo que cada elemento tem de similar em relação a outros pertencentes do grupo.

A idéia básica é que elementos que componham um mesmo cluster devem apresentar alta similaridade (i.e., sejam elementos bem parecidos, seguem um padrão similar), mas devem ser muito dissimilares de objetos de outros clusters. Em outras palavras, toda clusterização é feita com objetivo de maximizar a homogeneidade dentro de cada cluster e maximizar a heterogeneidade entre clusters.

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.

Uma definição formal do problema de Clusterização é encontrada em HRUSCHKA & EBECKEN (2001). Considerando um conjunto de *n* objetos $X = \{X_1, X_2, ..., X_n\}$ onde cada $X_i \in \mathbb{R}^p$é um vetor de p medidas reais que dimensionam as características do objeto, estes devem ser clusterizados em k clusters disjuntos $C = \{C_1, C_2, ..., C_k\}$, de forma que tenhamos as seguintes condições respeitadas:

1. $C_1 \cup C_2 \cup ... \cup C_k = X$;
2. $C_i \neq \emptyset, \forall i, 1 \leq i \leq k$;
3. $C_i \cap C_j = \emptyset, \forall i \neq j, 1 \leq i \leq k, 1 \leq j \leq k$.

Enfatiza-se que, por essas condições, um objeto não pode pertencer a mais de um cluster (grupos disjuntos) e que cada cluster tem que ter ao menos um objeto. COLE (1998) ainda acrescenta que o valor de *k* geralmente é desconhecido. Se *k* é conhecido, o problema é referido como o problema de k-Clusterização.

4.1. Aplicações

A Clusterização pode ser empregada quando objetivo é reduzir o número de objetos, para um número de subgrupos característicos, levando as observações a ser consideradas como membros de um grupo e perfiladas segundo características gerais que rotulam distintamente este grupo, ou também quando o pesquisador deseja formular hipóteses sobre a natureza dos dados ou examinar hipóteses pré-estabelecidas. Se uma determinada estrutura pode ser previamente definida para um certo grupo de objetos, o

resultado da análise de clusters pode ser utilizado para fins de comparação e validação daquela estrutura inicial. Outra utilidade da clusterização é na identificação de relacionamentos entre as observações não identificados em outras técnicas. Entretanto, o uso mais tradicional da Clusterização tem sido para propósitos exploratórios e formação de uma taxonomia, uma classificação de objetos com base empírica.

A área de Clusterização tem desenvolvimento vigoroso. É exaustivamente difícil listar todas as aplicações que têm utilizado técnicas de agrupamento distintas, bem como a grande quantidade de algoritmos publicados.A popularização do uso e desenvovimento de métodos de clusterização ocorrem devido à grande quantidade de dados coletados nas diversas áreas do conhecimento e atividade que tornam a análise de cluster um tópico altamente atrativo em várias pesquisas na mineração de dados.

Como exemplos de áreas interessadas no problema de Clusterização, podemos citar: mineração de dados, estatística, engenharia, aprendizado de máquina, medicina, marketing, administração e biologia. São comuns aplicações relativas a reconhecimento de padrões, análise de dados, processamento de imagens, pesquisa de mercado, padrão de compra, especificações físicas e químicas de petróleos, análise de sintomas de doenças, características de seres vivos, funcionalidades de genes, a composição de solos, aspectos da personalidade de indivíduos, perfis de clientes, marketing, segmentação de imagens, agrupamento de documentos, tecnologia da informaçãogestão de força de trabalho e planejamento, estudos de dados de genomana biologia, dentre muitas outras. Enfim, as aplicações a clusterização são utilizadas para cumprir pelo menos um dos seguintes objetivos principais:

- Identificação da estrutura subjacente: para obter ‘insights’ sobre os dados, gerar hipóteses, detectar anomalias, e identificar características marcantes.
- Classificação Natural: identificar o grau de semelhança entre as formas ou organismos (filogenética).
- Compressão: como um método para a organização dos dados e resumindo-o através de protótipos do cluster.

4.2. Limitações

Encontrar o melhor agrupamento para um conjunto de objetos não é uma tarefa simples. Como HRUSCHKA & EBECKEN (2001) destacam, o problema de encontrar esse melhor agrupamento é NP-completo e não é computacionalmente possível encontrá-lo, a não ser que *n* (número de objetos) e *k* (número de clusters) sejam extremamente pequenos, visto que o número de partições distintas em que podemos dividir *n* objetos em *k* clusters aumenta aproximadamente como $\frac{k^n}{n!}$.

ANKERST *et al.* (1999) escrevem que existem três razões interconectadas para explicar porque a efetividade dos algoritmos de Clusterização é um problema. Primeiro, quase todos os algoritmos de Clusterização requerem valores para os parâmetros de entrada que são difíceis de determinar, especialmente para conjuntos de dados do mundo real contendo objetos com muitos atributos. Segundo, os algoritmos são muito sensíveis a estes valores de parâmetros, frequentemente produzindo partições muito diferentes do conjunto de dados mesmo para ajustes de parâmetros significativamente pouco diferentes. Terceiro, os conjuntos de dados reais de alta dimensão têm uma distribuição muito ampla que não pode ser revelada por um algoritmo de Clusterização usando somente um ajuste de parâmetro global.

4.3. Medidas de Similaridade

O interesse de Clusterização consiste em formar grupos de objetos onde os elementos dentro decada grupo têm que ser mais similares entre si do que em relação aos elementos de grupos distintos. Para tal é necessário quantificar a similaridade entre os objetos. Para os algoritmos de Clusterização poderem efetuar sua tarefa é necessário que eles utilizem estruturas de dados capazes de armazenar os objetos a serem processados ou as informações sobre as relações entre estes.

Algoritmos de Clusterização que trabalham com dados armazenados na memória principal, tipicamente, usam uma das seguintes estruturas de dados no seu processamento.

- Matriz de dados: as linhas representam cada um dos objetos a serem clusterizados e as colunas, os atributos ou características de cada objeto. Considerando *n* objetos cada qual com *p* atributos, tem-se uma matriz *n* x *p* como abaixo:

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2p} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3p} \\ x_{41} & x_{42} & x_{43} & \cdots & x_{4p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{np} \end{bmatrix} \tag{40}$$

- Matriz de dissimilaridade: cada elemento da matriz representa a distância entre pares de objetos. Aqui, considerando *n* objetos a serem clusterizados têm-se uma matriz quadrada D_{nxn} como a que segue:

$$\begin{bmatrix} 0 & d(1,2) & d(1,3) & & d(1,n) \\ d(2,1) & 0 & d(2,3) & \ldots & d(2,n) \\ d(3,1) & d(3,2) & 0 & & d(3,n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d(n,1) & d(n,2) & d(n,3) & \cdots & 0 \end{bmatrix} = D \tag{41}$$

Com *d*(*i*, *j*) representando a distância ou dissimilaridade entre o objeto *i* e o *j*.

As medidas de similaridade são números positivos que exprimem a "distância" entre dois objetos. Quanto menor o valor de *d(i, j)*, mais semelhantes serão os objetos e de acordo com um critério, ficarão no mesmo cluster. De outro modo, quanto maior a "distância", menos similares serão os objetos e, em consequência, eles deverão estar em grupos distintos, pelo mesmo critério. Quando um algoritmo que trabalha com matrizes de dissimilaridade recebe uma matriz de dados, ele primeiro transforma-a em uma matriz de dissimilaridade antes de iniciar suas etapas de clusterização (HAN & KAMBER, 2001).

COLE (1998) resume que para clusterizar objetos de acordo com sua similaridade, deve-se definir uma medida de quão próximos dois objetos estão, ou quão bem seus valores se comparam. Uma pequena distância entre os objetos deve indicar uma alta similaridade. Assim, uma medida de distância pode ser usada para quantificar a similaridade. Uma função de distância deve ser definida de tal forma que obedeça as seguintes propriedades;

- Positividade: $d(i,j) \geq 0$;

- Simetria: $d(i,j) = d(j,i)$;
- Reflexiva: $d(i,j) = 0 \Leftrightarrow i = j$;
- Desigualdade Triangular: $\mathrm{d}(i,j) \leq d(i,h) + d(h,j)$.

Não há uma medida de similaridade que sirva para todos os tipos de variáveis que podem existir numa base de dados. As medidas que são normalmente usadas para computar as similaridades de objetos descritos por tais variáveis são: Euclidiana, Manhattan, Minkowski e Mahalanobis.

A unidade da medida usada pode afetar a análise. Para evitar isso, sugere-se normalizar os dados antes da clusterização. A normalização é efetuada para cada variável f (cada atributo dos objetos) da seguinte forma:

1. Calcular a média do desvio absoluto, *sf*:

$$s_f = \frac{1}{n}\left(\left|x_{1f} - m_f\right| + \left|x_{2f} - m_f\right| + \cdots + \left|x_{nf} - m_f\right|\right) \quad (42)$$

Os valores x_{1f} a x_{nf} são os valores do atributo f para os n objetos a serem clusterizados e m_f, é o valor médio do atributo *f*, isto é:

$$m_f = \frac{1}{n}\left(x_{1f} + x_{2f} + \cdots + x_{nf}\right) \quad (43)$$

O valor do i-ésimo objeto da variável f normalizado será dado por

$$z_{if} = \frac{x_{if} - m_f}{s_f} \quad (44)$$

Com ou sem a normalização, a similaridade entre os objetos descritos por variáveis escaladas em intervalos são computadas baseado na distância entre cada par de objetos. COLE (1998), HAN e KAMBER (2001) destacam que a mais utilizada é a distância Euclidiana, que é a distância em linha direta entre os dois pontos que representam os objetos.

Considerando objetos com p atributos, a distância Euclidiana é dada por:

$$\mathrm{d(i,j)} = \sqrt{\left|\mathrm{x_{i1}} - \mathrm{x_{j1}}\right|^2 + \left|\mathrm{x_{i2}} - \mathrm{x_{j2}}\right|^2 + \cdots + \left|\mathrm{x_{jp}} - \mathrm{x_{ip}}\right|^2} \quad (45)$$

A segunda medida de distância mais usada é a Manhattan ou “city-block”, que é a soma dos módulos das diferenças entre todos os atributos dos dois objetos em questão, ou seja:

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots |x_{ip} - x_{jp}| \quad (46)$$

Essa medida de similaridade é mais facilmente calculada do que a anterior, mas ela pode não ser adequada se os atributos estão correlacionados, pois não há garantia da qualidade dos resultados obtidos (COLE, 1998).

A distância Minkowski é a generalização das distâncias anteriores. Ela é representada por:

$$d(i,j) = \left(|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \cdots + |x_{ip} - x_{jp}|^q\right)^{\frac{1}{q}} \quad (47)$$

onde q é um inteiro positivo que no caso da distância Euclidiana é igual a 2 e no da city-block é igual a 1.

Em algumas análises de Clusterização há interesse em se aumentar a importância de um atributo ou conjunto de atributos em relação aos demais, nesse caso, atribui-se pesos a cada um dos atributos. Isso pode ser feito para todas essas medidas de distâncias. No caso da distância Minkowski, temos:

$$d(i,j) = \left(w_1|x_{i1} - x_{j1}|^q + w_2|x_{i2} - x_{j2}|^q + \cdots + w_p|x_{ip} - x_{jp}|^q\right)^{\frac{1}{q}} \quad (48)$$

Outra distância abordada na literatura e muito usada em clusterização é a Distância de Mahalanobis, que é dada por:

$$d(i,j) = (x_i - x_j)^T S^{-1} (x_i - x_j) \quad (49)$$

Onde S é a matriz de covariâncias entre grupos, calculada com todos os objetos. Se a matriz de covariância é a matriz identidade, a distância de Mahalanobis coincide com a distância euclidiana. Se a matriz de covariância é diagonal, então a medida de distância resultante é chamada distância euclidiana normalizada.

A distância de Mahalanobis é amplamente utilizada em análise de clusters. Justifica-se pela seguinte explicação: considere-se o problema de estimar a probabilidade de um ponto de teste no espaço euclidiano N-dimensional pertencer a um cluster que tem pontos amostrais que pertencem a esse clustet. Um primeiro passo poderia ser a determinação da média do centro de massa dos pontos amostrais. Intuitivamente, quanto mais próximo estiver o ponto em questão deste centro de massa, mais provável é que pertença ao conjunto. Quanto mais distante esteja, mais provável é que o ponto não deva ser classificado como pertencente ao conjunto. Todavia, precisa-se também de determinar a dimensão do conjunto. A distância de Mahalanobis é invariante a qualquer transformação linear não singular e tende a formar clusters hiperelípticos.

A distância de Mahalanobis tem em conta a variabilidade. Em vez de tratar todos os valores de igual modo, quando calcula a distância ao ponto central, pondera-os pela diferença à amplitude de variação na direção do ponto de teste. A fronteira de Mahalanobis torna-se assim clara. Para que se possa usar a distância de Mahalanobis na classificação de um ponto de teste quanto à pertença a um entre K clusters, convirá inicialmente determinar a matriz de covariância de cada classe, habitualmente baseando-se em amostras que se saibam pertencer a cada uma dessas classes. Então, dada uma amostra para teste, calcula-se a distância de Mahalanobis a cada uma dessas classes, e classifica-se o ponto de teste como pertencente à classe com a qual a distância de Mahalanobis seja a menor de todas. Usando a interpretação probabilística acima referida, isto é equivalente à seleção da classe que apresente a máxima verosimilhança.

4.4. Métodos de Clusterização

Um método ideal de Clusterização deveria atender aos seguintes requisitos (AGRAWAL *et al.* (1998); ESTER *et* al. (1996); NG & HAN (1994); HAN & KAMBER (2001)):

a) descobrir clusters com forma arbitrária - a forma dos clusters, considerando o espaço Euclidiano, pode ser esférica, linear, alongada, elíptica, cilíndrica, espiralada, etc.

b) identificar clusters de tamanhos variados;

c) aceitar os diversos tipos de variáveis possíveis - os métodos têm que ser capazes de lidar com as variáveis dos tipos: escaladas em intervalos, binárias, nominais (categóricas), ordinais, escaladas em proporção, ou ainda combinações desses tipos de variáveis;

d) ser insensível a ordem de apresentação dos objetos - um mesmo conjunto de objetos quando apresentado com diferentes ordenamentos deve fornecer os mesmos resultados

e) trabalhar com objetos com qualquer número de atributos (dimensões) – os olhos humanos são bons para julgar a qualidade de Clusterização com até três dimensões, os métodos devem manejar, com eficiência, objetos com altas dimensões e fornecer resultados inteligíveis onde a visualização humana é impossível.

f) ser escalável para lidar com qualquer quantidade de objetos – uma base de dados de grande porte pode conter milhões de objetos. Os métodos devem ser rápidos e escalonáveis com o número de dimensões e com a quantidade de objetos a ser clusterizado;

g) fornecer resultados interpretáveis e utilizáveis - as descrições dos clusters devem ser facilmente assimiladas, os usuários esperam que os resultados das Clusterizações sejam interpretáveis, compreensíveis e utilizáveis, é importante ter representações simples;

h) ser robusto na presença de ruídos - a maioria das bases de dados do mundo real contém ruídos ou dados perdidos, desconhecidos ou errados, a existência deles não deve afetar a qualidade dos clusters obtidos;

i) exigir o mínimo de conhecimento para determinar os parâmetros de entrada; os valores apropriados, são frequentemente, desconhecidos e difíceis de determinar, especialmente, para conjuntos de objetos de alta dimensionalidade e de grande número de objetos. Em alguns métodos, os resultados da Clusterização são bastante sensíveis aos parâmetros de entrada;

j) aceitar restrições - aplicações do mundo real podem necessitar agrupar objetos de acordo com vários tipos de restrições, os métodos devem encontrar grupos de dados com comportamento que satisfaça as restrições especificadas;

k) encontrar o número adequado de clusters - encontrar o número natural de clusters de um conjunto de objetos é uma tarefa difícil. Muitos métodos precisam de um valor de referência.

Entretanto, como dito por AGRAWAL *et al.* (1998), nenhuma técnica de Clusterização atende a todos estes pontos adequadamente, embora um trabalho considerável tenha sido feito para atender a cada ponto separadamente. Assim, há métodos apropriados para grandes quantidades de objetos e outros para pequenas quantidades; métodos em que o número de clusters tem que ser fornecido pelo usuário e outros em que não há essa exigência; métodos mais adequados a clusters de forma esférica ou convexa e outros que a forma do cluster não é relevante; métodos capazes de identificar clusters que tenham tamanhos diversos e outros que necessitam que os clusters tenham tamanhos semelhantes; métodos para dados categóricos; métodos que sofrem a influência de "ruídos" e outros insensíveis a estes; métodos para dados espaciais; métodos para espaços com elevado número de dimensões, etc.

Uma classificação geral dos algoritmos de Clusterização divide os algoritmos em dez tipos principais:

- Métodos Hierárquicos;
- Métodos Particionais;
- Métodos Baseados em Densidade;
- Métodos Baseados em Grade;
- Métodos Baseados em Modelos;
- Métodos Baseados em Redes Neurais;
- Métodos Baseados em Lógica Fuzzy;
- Métodos Baseados em Kernel;
- Métodos Baseados em Grafos;
- Métodos Baseados em Computação Evolucionária.

Os métodos mais tradicionais de Clusterização são os métodos Particionais e os métodos Hierárquicos. Como destacado por HAN & KAMBER (2001), alguns algoritmos de Clusterização integram as idéias de vários outros, então, algumas vezes, é difícil classificar um dado algoritmo como unicamente pertencendo a somente uma categoria de método de Clusterização. Além do que, algumas aplicações podem ter critérios de Clusterização que requerem a integração das várias técnicas de Clusterização acima. Nas subseções seguintes, cada um dos cinco métodos de Clusterização é caracterizado.

1.1.1. Métodos Hierárquicos

Algoritmos de clusterização baseados no método hierárquico (HC) organizam um conjunto de dados em uma estrutura hierárquica de acordo com a proximidade entre os indivíduos. Os resultados de um algoritmo HC são normalmente mostrados como uma árvore binária ou dendograma, que é uma árvore que iterativamente divide a base de dados em subconjuntos menores. A raiz do dendograma representa o conjunto de dados inteiro e os nós folhas representam os indivíduos. O resultado da *clusterização* pode ser obtido cortando-se o dendograma em diferentes níveis de acordo com o numero de cluster k desejado. Esta forma de representação fornece descrições informativas e visualização para as estruturas de grupos em potencial, especialmente quando há realmente relações hierárquicas nos dados como, por exemplo, dados de pesquisa sobre evolução de espécies. Em tais hierarquias, cada nó da árvore representa um cluster da base de dados. O *dendrograma* pode ser criado de duas formas:

1. Abordagem aglomerativa (bottom-up) parte-se das folhas superiores para a raiz. Inicia-se considerando cada objeto como sendo um cluster, totalizando *n* clusters. Em cada etapa, calcula-se a distância entre cada par de clusters. Estas distâncias são geralmente, armazenadas em uma matriz de dissimilaridade simétrica. Daí, escolhe-se 2 clusters com a distância mínima e junta-os. A seguir, atualizamos a matriz de distâncias. Este processo continua até que todos os objetos estejam em um único cluster (o nível mais alto da hierarquia), ou até que uma condição de

término ocorra (AGRAWAL *et al.*, 1998; NG & HAN, 1994; HAN & KAMBER, 2001);

2. Abordagem divisiva (top-down) parte-se da raiz para as folhas. Nesta abordagem o processo é o inverso da abordagem bottom-up por começar com todos os objetos em um único cluster. Em cada etapa, um cluster é escolhido e dividido em dois clusters menores. Este processo continua até que se tenham *n* clusters ou até que uma condição de término, por exemplo, o número de clusters k desejado aconteça.

Os métodos aglomerativos são mais populares do que os métodos divisivos. ZHANG et al. (1996) dizem que os métodos hierárquicos não tentam encontrar os melhores clusters, mas manter junto o par mais próximo (ou separar o par mais distante) de objetos para formar clusters. E também salientam que a melhor estimativa para a complexidade de um algoritmo prático por método hierárquico é $O(n^2)$ o que o torna incapaz de ser eficiente para *n* grande.

1.1.2. Métodos Particionais

Os algoritmos particionais dividem a base de dados em *k*-grupos, onde o número *k* é dado pelo usuário, como ESTER *et al.* (1996) lembram que este é um ponto negativo do método pois esse domínio de conhecimento não é disponível para muitas aplicações.

Inicialmente, o algoritmo escolhe *k* objetos como sendo os centros dos *k* clusters. Os objetos são divididos entre os *k* clusters de acordo com a medida de similaridade adotada, de modo que cada objeto fique no cluster que forneça o menor valor de distância entre o objeto e o centro do mesmo. Então, o algoritmo utiliza uma estratégia iterativa de controle para determinar que objetos devem mudar de cluster, de forma que a função objetivo usada seja otimizada.

Após a divisão inicial, há duas possibilidades na escolha do "elemento" que vai representar o centro do cluster, e que será a referência para o cálculo da medida de

similaridade. Ou utiliza-se a média dos objetos que pertencem ao cluster em questão, também chamada de centro de gravidade do cluster (esta é a abordagem conhecida como *k-means)*; ou escolhe-se como representante o objeto que se encontra mais próximo ao centro de gravidade do cluster (abordagem é conhecida como *k-medoids),* sendo o elemento mais próximo ao centro chamado de *medoid*.

O k-means é o mais popular e mais simples algoritmo particional. K-means foi descoberto independentemente em diferentes campos científicos, primeiramente por STEINHAUS (1956), LLOYD (1982) (na verdade, proposto em 1957, mas publicado somente em 1982), BALL & HALL (1965) e MACQUEEN (1967) e mesmo tendo sido proposto há mais de 50 anos, ainda é um dos algoritmos mais utilizados para clusterização devido à facilidade de implementação, simplicidade, eficiência e sucesso empírico e possui várias extensões desenvolvidas em várias formas (JAIN, 2009).

A função objetivo mais utilizada para espaços métricos nos métodos particionais é o erro quadrático, dado por):

$$E = \sum_{j=1}^{k} \sum_{x \in C_i} \left\| p - m_i \right\|^2, \quad para \ \ k \in (1, n). \qquad (50)$$

Na equação, E é a soma do erro quadrado para todos os objetos na base de dados, p é o ponto no espaço representando um dado objeto, e m_i é o representante do cluster C_i. Tanto p quanto m_i são multidimensionais. Essa função objetivo dividida por n representa a distância média de cada objeto ao seu respectivo representante (ESTER *et al.*, 1998) e também é conhecida como critério do erro médio quadrado. Os algoritmos terminam quando não existem atribuições possíveis capazes de melhorar esta função objetivo (COLE, 1998).

Ao contrário dos métodos hierárquicos, que produzem uma série de agrupamentos relacionados, métodos particionais produzem agrupamentos simples. ANKERST *et al.* (1999) destacam que esses algoritmos são efetivos se o número de clusters *k* puder ser razoavelmente estimado, se os clusters são de forma convexa e possuem tamanho e densidade similares. GUHA *et al.* (1998) colocam que os métodos particionais tentam fazer os *k* clusters tão compactos e separados quanto possível, e que trabalham bem quando os clusters são compactos, densos e bastante separados uns dos outros, mas não são tão eficientes quando existem grandes diferenças nos tamanhos e geometrias dos

diferentes clusters. HAN & KAMBER (2001) observam que os mais bem conhecidos e geralmente usados métodos de particionamento são o k-means, o K-medoids, e suas variações.

1.1.3. Métodos Baseados em Densidade

A maioria dos métodos Particionais clusteriza objetos com base na distância entre eles. Tais métodos podem encontrar dificuldades para descobrir clusters de formas arbitrárias. Nos métodos de Clusterização baseados em Densidade, clusters são definidos como regiões densas, separadas por regiões menos densas que representam os ruídos. As regiões densas podem ter uma forma arbitrária e os pontos dentro de uma região podem também estar distribuídos arbitrariamente e, por isso, os métodos baseados em densidade são adequados para descobrir clusters com forma arbitrária, tais como elíptica, cilíndrica, espiralada, etc. até os completamente cercados por outro "cluster" e são especialistas em identificar e filtrar ruídos, HAN & KAMBER (2001). Os métodos baseados em densidade diferem-se pela forma com que crescem os clusters: uns determinam os clusters de acordo com a densidade da vizinhança dos objetos, outros, trabalham de acordo com alguma função de densidade.

Para entender a idéia dos métodos baseados em densidade, ESTER *et al.* (1996) observam que ao visualizar conjuntos de objetos tais como os da Figura 4.1, a seguir, pode-se, facilmente, e de forma não ambígua, detectar clusters circulares no conjunto 1, clusters de formatos arbitrários no conjunto 2 e clusters de objetos e óbvios ruídos não pertencentes a qualquer dos clusters no conjunto 3. Tem-se este reconhecimento automático porque o cérebro humano reconhece visualmente, neste caso bidimensional, que dentro de cada cluster tem-se uma densidade de objetos típica que é consideravelmente maior do que fora dos clusters. Além disso, a densidade de áreas de ruído é menor do que a densidade em qualquer dos clusters. O cérebro humano reconhece os clusters e ruídos nos exemplos mostrados na Figura 4.1 usando involuntariamente o conceito de grupos formados por densidade.

Um método baseado em densidade clusteriza objetos baseado nesta noção de densidade e capta este comportamento, observado visualmente de forma óbvia nestes exemplos bidimensionais, também em conjuntos de dados de pontos de maiores dimensões, ou seja, em qualquer espaço p-dimensional, $p > 2$, onde a visualização humana não é capaz de compreender a formação dos clusters, com tal facilidade.

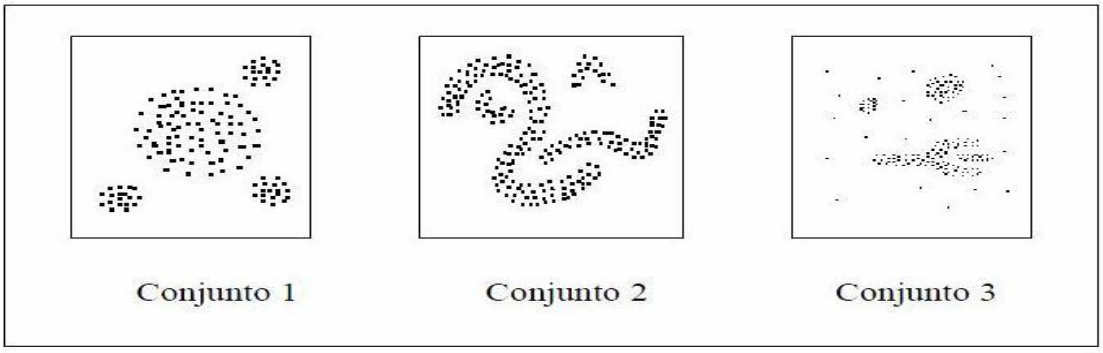

Figura 4.1. Conjunto com clusters globulares (conjunto 1), não globulares (conjuntos 2 e 3) e com ruídos (conjunto 3). Fonte: ESTER et al. (1996).

1.1.4. Métodos Baseados em Grade

Os métodos de Clusterização baseados em grade usam uma estrutura de dados em grade de multiresolução. Eles discretizam o espaço de objetos em um número finito de células que formam uma estrutura de grade na qual todas as operações de Clusterização são efetuadas. A principal vantagem desta abordagem é seu tempo de processamento rápido, que é tipicamente independente do número de objetos de dados, contudo dependente, somente, do número de células em cada dimensão no espaço discretizado, HAN & KAMBER, 2001.

1.1.5. Métodos Baseados em Modelos

São métodos que usam um modelo de referência para cada cluster. Eles tentam otimizar a curva entre os objetos dados e algum modelo matemático.Um algoritmo baseado em modelo pode descobrir clusters construindo uma função de densidade que reflete a distribuição espacial dos pontos de dados. Ele também conduz a um modo de determinar automaticamente o número de clusters baseado na estatística padrão,

identificando ruídos no relatório e assim produzindo métodos de Clusterização robustos. Tais modelos são, frequentemente, baseados na suposição que os dados são gerados por uma mistura de distribuições de probabilidades.

Ao contrário dos métodos de Clusterização convencionais, que primariamente identificam grupos de objetos, os métodos de Clusterização Baseados em Modelos, também chamados de Métodos de Clusterização Conceitual, realizam uma etapa adicional para encontrar descrições características para cada grupo, onde cada grupo representa um conceito ou classe. Sendo assim, a qualidade de Clusterização não é unicamente uma função dos objetos individuais. Antes, o método incorpora fatores tais como a generalidade e a simplicidade das descrições conceituais derivadas. Muitos métodos de Clusterização adotam uma abordagem estatística que usa medidas de probabilidade na determinação dos conceitos ou clusters. Nestes casos,
descrições probabilísticas são usadas para representar cada conceito derivado. Mais detalhes sobre este tipo de clusterização estão em HAN & KAMBER (2001).

1.1.6. Métodos Baseados em Redes Neurais Artificiais

As redes neurais artificiais (RNA's) são apropriadas para tarefas de percepção como o reconhecimento, classificação e autoassociação de padrões. Para o projeto de classificadores, a abordagem por rede neural para Clusterização tende a representar cada cluster como um exemplar. Um exemplar serve de protótipo do cluster e não necessariamente corresponde a um exemplo de dado particular ou objeto. Novos objetos podem ser distribuídos para clusters cujo exemplar é mais similar baseado em alguma medida de distância. Os atributos de um objeto atribuído a um cluster podem ser preditos dos atributos do exemplar do cluster.

As RNA's apresentam uma série de vantagens na clusterização como robustez ao ruído, capacidade de generalização, aprendizado adaptativo a partir de exemplos e processamento paralelo. Os métodos de clusterização baseados em redes neurais ou tem suas raízes no método de clusterização ART (Teoria Ressonante Adaptativa) proposto

em 1988 por CARPENTER & GROSSBERG (1988), ou nos Mapas Auto Organizáveis de Kohonen, KOHONEN (1989). As redes neurais usadas em ART implementam de forma complexa a idéia do algoritmo convencional de agrupamentos 'leader-follower' proposto em SPÄTH (1980), que é extremamente sensível à sequência de apresentação dos dados e ao parâmetro que controla o raio máximo em um agrupamento existente. Por isso a abordagem de clusterização mais difundida e explorada baseada em redes neurais é feita pelos mapas de Kohonen.

1.1.7. Métodos Baseados em Lógica *Fuzzy*

Em abordagens tradicionais de Clusterização cada objeto pertence, ao final, a um e somente um cluster. Portanto, os clusters nesses tipos de abordagens são disjuntos. Métodos com esta característica são métodos de Clusterização 'hard'.

Os métodos de clusterização baseados em Lógica Fuzzy são métodos não 'hard', que permitem associar um indivíduo a todos os clusters usando uma função de pertinência (ZADEH, 1965). A restrição adotada nesta metodologia é que a soma dos graus de pertinência de um indivíduo aos clusters seja igual a 1. Em um algoritmo de clusterização fuzzy, cada cluster é um conjunto fuzzy de todos os indivíduos. O conceito de conjuntos fuzzy oferece a vantagem de expressar o tipo de situação em que um indivíduo compartilha similaridade com vários grupos, o algoritmo assim associa cada indivíduo parcialmente a todos os grupos.

1.1.8. Métodos Baseados em *Kernel*

Método *Kernel* é uma técnica desenvolvida especialmente para problemas não linearmente separáveis serem resolvidos de forma 'mais elegante'. Algoritmos Kernel usam do espaço de características para permitir uma separação não-linear no espaço de entrada. Clusterização Baseada em Kernel usam esta abordagem e realizam o agrupamento implicitamente por um método Kernel, que executa um mapeamento não

linear apropriado dos dados de entrada para um espaço de características de alta dimensão, ao substituir o produto interno entre as variáveis não-lineares por um determinado Kernel apropriado denominado Mercer Kernel , como descreve CRISTIANINI & SHAWE-TAYLOR (2000). Sendo assim, clusterização baseada em métodos Kernel são capazes de produzir uma separação não linear entre os hiperespaços dos clusters, ao contrário dos algoritmos tradicionais que produzem por partes fronteiras lineares entre os dados, JAIN et al. (1999).

Clusterização baseada em Kernel tem muitas vantagens: é possível obter um hiperplano linearmente separável no espaço de alta dimensão, ou de característica infinito; pode formar clusters de formas arbitrárias, diferentes, hiperelipsoidais e hiperesféricas, tem a capacidade de lidar com o ruído e outliers e não há nenhuma exigência de conhecimento prévio para determinar a estrutura topológica do sistema.

1.1.9. Métodos Baseados em Grafos

Os algoritmos de clusterização baseados em grafos buscam representar um conjunto de dados em um grafo, onde cada vértice representa um elemento do conjunto de dados e a existência de uma aresta conectando dois vértices é feita com base na proximidade entre os dois dados. A maneira mais simples de estabelecer as ligações entre os vértices é conectar cada vértice aos vértices restantes, onde o peso indica a similaridade entre os dois dados e um cluster é definido como um subgrafo do grafo inicial. Para tal, adota-se uma medida de similaridade no processo de agrupamento, o que pode fazer com que o algoritmo apresente alguma dificuldade em determinar clusters de formas variadas. Os algoritmos de clusterização baseados em grafos são fortemente relacionados com os algoritmos hierárquicos e particionais. Isso significa que o resultado obtido pode ser uma partição ou uma hierarquia de partições. O trabalho de SCHAEFFER (2007) oferece uma boa revisão sobre diferentes métodos de clusterização baseados em grafos.

1.1.10. Métodos Baseados em Computação Evolucionária

A Computação Evolucionária compreende um conjunto de técnicas de busca e otimização baseados em mecanismos da evolução biológica, tais como reprodução, mutação, recombinação e seleção natural e estão sendo utilizados amplamente pela comunidade de inteligência artificial para obter modelos de inteligência computacional. Em tais abordagens, o conjunto de dados representa a população sob evolução e seu comportamento é simulado através de repetidas operações associadas aos princípios de mutações genéticas e de seleção natural, comuns na evolução biológica.

Sub-grupos dos algoritmos evolutivos foram criados de acordo com a aplicação para a qual são destinados. O mais popular é o algoritmo genético, no qual busca a solução através de operações, como mutação, sobre cadeias de números, geralmemente binários e conforme foi colocado por COLE (1998), os Algoritmos Genéticos têm sido feitos para serem aplicados ao problema de Clusterização por adaptarem a representação, a função de adequação, e desenvolver operadores evolucionários adequados para o problema de agrupamento. Como foi comentado na seção 2.2., o problema de Clusterização é NP-completo, o número de possíveis combinações em que se pode particionar os *n* objetos em *k* clusters cresce rapidamente e um algoritmo com um potencial para fazer buscas em espaços de solução grandes efetivamente se faz necessário. Os Algoritmos Genéticos têm habilidade em cobrir um subconjunto grande do espaço de busca, são, de maneira geral, efetivos em problemas de otimização global NP-completos e eles podem prover boas soluções sub-ótimas em tempo razoável.

Uma abordagem interessante é a descrita em HRUSCHKA & EBECKEN (2001), onde os autores descrevem um Algoritmo Genético para análise de cluster. Os autores adotam um esquema de codificação simples que conduz a cromossomos de comprimento constante. A função objetivo maximiza a homogeneidade dentro de cada cluster e a heterogeneidade entre cluster e encontram o número de clusters de acordo com a largura de silhueta média. Sendo assim, a função objetivo maximiza a homogeneidade dentro de cada cluster, a heterogeneidade entre clusters. Os autores defendem que um Algoritmo Genético de Clusterização pode prover um caminho para

encontrar a Clusterização correta, encontrar o número correto de clusters, com a vantagem de não precisar de parâmetros de entrada, e atende a outros requisitos desejáveis para Clusterização tais como a insensibilidade à ordem de entrada dos dados, CARLANTONIO (2001).

4.5. Histórico dos Métodos de Clusterização

De acordo com BAILEY (1975), estudos de Clusterização se originaram na Antropologia, a partir do trabalho de DRIVER & KROEBER (1932) e na Psicologia, a partir dos trabalhos de ZUBIN (1938) e TRYON (1939). No entanto, durante muito tempo as diferentes técnicas de Clusterização ficaram restritas a um grupo reduzido de pesquisadores devido a sua complexidade matemática. O desenvolvimento da tecnologia computacional é o elemento de maior valia para explicar a propagação da técnica entre os diferentes ramos do conhecimento e o desenvolvimento de novos métodos. Pela quantidade e diversificação de métodos e algoritmos propostos é extremamente difícil rever todas as abordagens publicadas na área de Clusterização. Nesta Seção são citadas as principais abordagens no tempo, 125 distintos métodos de clusterização, lançados de 1948 até 2014.

O primeiro registro publicado de um método de Clusterização se deu em SORENSEN (1948), onde o autor definiu o método **Hierárquico Aglomerativo de Ligação Completa** e depois deste, o mais popular método de clusterização, o **k-means,** foi lançado no trabalho STEINHAUS (1956). Somente nove anos depois do lançamento do Método Hierárquico de Ligação Completa, foi lançado o método **Hierárquico de Ligação Simples**, por SNEATH (1957).

WARD (1963) apresentou um novo critério aplicado na análise de clusterização hierárquico chamado método de **clusterização de Ward**, que minimizava o total de variância intra-cluster. Logo depois foram lançados dois métodos variantes do K-means: o **FORGY,** em FORGY (1965) e o **ISODATA**, em BALL & HALL (1965), que faz uma realocação iterativa e auto-organizável dos dados.

Em 1972, o conceito de **co-clusterização de Hartigan** (ou clusterização bi-dimensional) foi introduzido no trabalho HARTIGAN (1972). Também neste ano, DIDAY (1972) apresentou a **Clusterização por Nuvens Dinâmicas**, um método não hierárquico que define as formas pesadas de um conjunto de curvas e tipologias. O primeiro método para considerar a presença de ruído na base de dados foi o **SNN** (Shared Nearest Neigbor), proposto por JARVIS & PATRICK (1973). Também neste ano surgiu o **Fuzzy C-means** propostos por DUNN (1973), que é uma extensão do K-means baseada em lógica fuzzy. Outra abordagem de clusterização baseada em lógica fuzzy foi o método **Backer** proposto em BACKER (1978). Dois anos depois, uma variante do **K-means, utilizando a distância Itakura-Saito** foi proposta para quantização vetorial no processamento da fala em LINDE et al. (1980). Depois, BEZDEK (1981) desenvolveu o **Método Fuzzy C-means Melhorado**.

Em 1983, uma nova abordagem de Clusterização Baseada em Pesquisa Combinatória foi apresentado por KIRKPATRICK et al. (1983) chamada de **Clustering SA (Simulated Annealing)**. Em 1987, FISHER (1987) inventou o **COBWEB**, um algoritmo de clusterização hierárquico popular, baseado também em modelos, que faz uma única passagem pelos dados disponíveis e organiza-os em uma árvore de classificação de forma incremental.

JAIN (1988) foi o primeiro a explorar uma abordagem baseada em densidade para identificar clusters em dados espaciais. No método **JAIN** o conjunto de dados é particionado para um número de células que não se sobrepõem e os histogramas são construídos. Células com contagem relativamente alta frequência de pontos são os centros dos grupos potenciais e as fronteiras entre grupos caem nos "vales" do histograma. É um método que identifica clusters de qualquer forma, mas o a memória requisitada e o tempo de execução para o armazenamento e a pesquisa nos histogramas multidimensionais podem ser enormes.

A primeira proposta de clusterização baseada em redes neurais artificiais também foi lançada em 1988, a **ART (Adaptive Resonance Theory)**, por CARPENTER & GROSSBERG (1988). Um ano depois, outro método baseado na técnica de rede neural artificial foi proposto por KOHONEN (1989), o **SOM (Mapas Auto Organizáveis**). Em 1989, outro Método Baseado em Pesquisa Combinatória, o método **TS (Tabu**

Search), foi apresentado por GLOVER (1989). KAUFFMAN & ROUSSEEUW (1989) apresentaram cinco novos métodos: dois métodos de clusterização particional, **PAM (Partitional Around Medoids)** e **CLARA (Clustering LARge Aplications)**; dois métodos hierárquicos divisivos **DIANA (Divisive ANAlysis)** e **MONA (MONotetic Analysis)** e um método hierárquico aglomerativo, o **AGNES** (AGlomerative NESting). Também neste ano FISHER et al. (1989) desenvolveram o **CLASSIT**, um algorítimo hierárquico incremental, que é uma versão do COBWEB para atributos numéricos.

Clusterização Baseada em Grafos ganhou a primeira abordagem no algoritmo **Ratio Cut** de HAGEN & KAHNG (1992). Em YANG (1993) o autor fez uma combinação de clusterização fuzzy, algoritmo baseado em relação fuzzy, e a regra do k-ésimo vizinho mais próximo e desenvolveu o **FCS- Fuzzy C-Sheels**. Uma abordagem fuzzy para lidar com valores atípicos foi introduzida no **PCM- Probabilístico C-Means** por KRISHNAPURAM & KELLER (1993), neste método o efeito do ruído e de outliers é diminuído com a consideração de tipicidade.

O método de clusterização **CLARANS** (Clustering Large Applications based on RANdomized Search), o primeiro método de clusterização concebido com a finalidade KDD - Knowledge Discovery in Databases foi apresentado por HAN & NG (1994). Também neste ano, YAGER & FILEV (1994) propuseram o método de **Clusterização de Montanha**, um algoritmo fuzzy a fim de estimar os centros de clusters. O algoritmo hierárquico divisivo **Ejcluster** foi apresentado por GARCÍA et al. (1994) na área de processamento de sinais. Ejcluster tem a vantagem de não necessitar de parâmetros de entrada e é muito eficaz na descoberta de agrupamentos não convexos. No entanto, o custo computacional de Ejcluster é $O(n^2)$ e o método não é robusto a presença de ruídos.

Uma estratégica taxa de aprendizagem adaptativa para o modo K-means on-line foi proposto por CHINRUNGRUENG & SÉQUIN (1995), o **Opt- K-Means** - Optimal adaptative K- means tem ajuste dinâmico da taxa de aprendizagem e pode ser ajustado, sem envolver quaisquer atividades do usuário. Para melhorar a eficiência, outro método de clusterização otimizada foi proposto, a **Clusterização Baseada em Amostragem por R-Tree**, em ESTER et al. (1995).

Estes autores também desenvolveram, em ESTER et al. (1996), o **DBSCAN (Density Based Spatial Clustering of Applications with Noise),** um método baseado

em densidades revolucionário no tratamento de ruído, semelhante ao algoritmo SNN Jarvis-Patrick, e que detecta clusters de formato arbitrários. **K-means com métrica de distância Mahalanobis** foi utilizada para detectar grupos hyperelipsoidais em MAO & JAIN (1996). Neste mesmo trabalho, os autores lançaram também o método o **HEC-(Gaussian Mixture Densities Decomposition)**, baseado em redes neurais artificiais que utiliza uma arquitetura de rede de duas camadas para estimar a distância de Mahalanobis regularizada. Ainda neste ano SCHIKUTA (1996) apresentou o **GridFile**, um método hierárquico e baseado em grades para grandes conjuntos de dados. ZHUANG et al. (1996) apresentaram o **BIRCH** (Balanced Iterative Reducing and Clustering Using Hierarchies) que faz uma redução iterativa otimizada na Clusterização Hierárquica e é baseado no recurso de cluster de árvore de fatores;e também o método probabilístico **GMDD** (Gaussian Mixture Densities Decomposition). CHEESEMAN & STUTZ (1996) propuseram um método particional probabilístico, muito usado na indústria, o **AutoClass**.

O método DBSCAN é muito explorado e tem diversas variações, pelo menos 19 algoritmos melhorados DBSCAN foram propostos, sendo o mais recente publicado em 2013. Basicamente, os investigadores tentam melhorar a complexidade técnica de DBSCAN e seu desempenho em densidades variadas utilizando o particionamento e abordagem híbrida. Nestes algoritmos, uma partição com diferentes parâmetros é formada e, em seguida, DBSCAN tradicional é aplicado. Na maioria dos casos a saída permanece a mesma ou é melhorada. Remoção de ruído e o efeito de alta dimensionalidade permanecem como no DBSCAN original.

ESTER et al. (1997) publicaram a primeira generalização para DBSCAN, o **GDBSCAN** (General DBSCAN). No mesmo ano, WANG et al . (1997) lançou o método **STING (STatistical INformation Grid)**, que usa uma quad-tree como estrutura contendo informação estatística adicional. Outro método probabilístico foi publicado em 1997 o **EM (Expectation- Maximization)**, por MCLACHLAN et al. (1997) que é uma generalização para os métodos k-means e fuzzy c-means. Em 1997, também foi lançado o **Modelo de Markov Escondido para Clusterização (HMM),** por MOZER et al. (1997), o método mais importante de Statistics Sequence Clustering, que ganhou a sua popularidade na aplicação de reconhecimento de voz.

O ano de 1998 foi marcado por uma grande produção na Área de Clusterização, uma vez que foram apresentados dez novos métodos de clusterização. O **K- modes** (que usa a simples medida do coeficiente de semelhança de lidar com atributos categóricos) e o **K- prothotypes** (através da definição de um conjunto de dissimilaridade medida, integra ainda as K-means e k- modes algoritmos para permitir casos de clusterização descritos por atributos mistos) foram propostos por HUANG (1998).

O **DENCLUE** (Density Clustering), outro método de clusterização baseado em densidade que detecta clusters de forma arbitrária, foi apresentado em HINNEBURG & KEIM (1998) (depois publicado em HINNEBURG & KEIM (2003)). O **CLIQUE**, (CLustering in QUEst) é um método de clusterização escalável baseado em densidade e grade que foi projetado para encontrar subespaços nos dados com alta densidade, por AGRAWAL et al. (1998); o **DBCLASD** (Distribution-Based Clustering Algorithm for Mining in Large Spatial Databases) introduzido por XU et al. (1998), aumenta de forma incremental um cluster inicial por seus pontos vizinhos, considerando que a distância do mais próximo vizinho oservada se adapte a uma distribuição esperada a distância. Ainda em 1998, o primeiro método utilizando wavelets, **WaveCluster,** foi proposto por SHEIKHOLESLAMI et al. (1998), sendo um método baseado em grade e densidade. O **K-means Kernel**, é uma proposta baseada em kernel para detectar clusters de formatos arbitrários, desenvolvida em SHOLKÖPF et al. (1998). O **CURE** (Clustering Using Representatives) é um método baseado em amostragem que é hierárquico aglomerativo, eficiente para grandes bancos de dados, proposto por GUHA et al. (1998). **Incremental DBSCAN** foi proposto por ESTER et al. (1998); e BRADLEY et al. (1998) também apresentou uma versão incremental, rápido e escalável de uma única passagem de K-means, o **Fast K-means**, que não necessita de todos os dados para se encaixar na memória, ao mesmo tempo, projetado para operar em uma única passagem sobre pontos de dados para melhorar a eficiência de clusterização de dados. Estes métodos incrementais estão em contraste com a maioria dos algoritmos de clusterização que exigem várias passagens sobre os pontos de dados antes de identificar os centros dos clusters.

O ano de 1999 foi também muito produtivo em relação à quantidade de métodos de clusterização publicados. Os primeiros métodos de clusterização baseados em

Computação Evolucionária surgiram em 1999: o **GGA** (Algoritmo Geneticamente Guiado), por HALL et al. (1999) e o **GKA** (Genetic k-means Algorithm), por KRISHNA & MURTYG. (1999). Outro algoritmo baseado em densidade também foi apresentado em 1999, o **OPTICS** (Ordering Points To Identify the Clustering Structure), por ANKERST et al. (1999). O inovador **CHAMALLEON,** baseado em teoria de grafos foi lançado em KARYPIS ET AL. (1999). O algoritmo **FDC** (Fast Density Clustering) foi apresentado por BO ZHOU et al. (1999) para clusterização com base em densidade definida por uma relação de equivalência sobre os objetos na base de dados. Também neste ano, o **PDBSCAN** (versão paralela de DBSCAN) foi apresentado por XU et al. (1999). O **MAFIA** (Merging of Adaptive Finite Intervals Approach), também surgiu em 1999 por HARASHA et al. (1999) . PELLEG & MOORE (1999) apresentaram o **Kd-tree k-means**, onde uma árvore KD é usada para identificar de forma eficiente os centros de clusters mais próximos para todos os pontos de dados, no k-means.

Estes autores também propuseram o **X-means** em PEELEG & MOORE (2000), que encontra automaticamente o número de clusters k, otimizando o Akaike Information Criterion (AIC) ou Bayesian Information Criterion (BIC). STEINBACH et al. (2000) propuseram uma versão hierárquica divisiva do K-means, chamado **K-means Divisivo**, ou bisecting k-means, que recursivamente particiona os dados em dois grupos em cada etapa. Outro método de clusterização baseado teoria dos grafos surgiu em 2000, o **HCS** (High Connected Subgraph - Subgrafos altamente conectados), por HARTUV et al. (2000). Também em 2000, surgiu na literatura o **ROCK**, um algoritmo de clusterização hierárquico robusto para atributos categóricos, utilizando o coeficiente de Jaccard para medir similaridade, proposto por GUHA et al. (2000). Um algoritmo eficiente baseado em aproximação gráfica de corte com restrição do tamanho do cluster (volume do cluster, ou soma dos pesos de ponta dentro de um cluster) chamado **Normalized Cut**, foi proposto em SHI & MALIK (2000). STEINBACH et al. (2000) publicaram um método de clusterização de dados por sumarização de dados chamado **Divide-and-Conquer.** Uma versão da DBSCAN também foi publicado em 2000, por ZHOU et al. (2000), o **SB-DBSCAN**. Neste trabalho, dois algoritmos DBSCAN baseados em amostragem (**SDBSCAN**) também foram desenvolvidos. Um algoritmo apresenta técnica de amostragem dentro do DBSCAN; e o outro usa procedimento de amostragem fora DBSCAN.

Em 2001, uma nova abordagem foi desenvolvida em clusterização: os métodos baseados em restrições, como o método **COD- Clustering com Obstructed Distances,** proposto por TUNG et al. (2001). E o segundo método kernel baseado desenvolvido, o **SVC- Support Vector Clustering**, foi publicado por BEN HUR et al. (2001). O método de **Entropia Mínima** apresentados em ROBERTS et al. (2001) assume que os dados são gerados utilizando um modelo de mistura e cada grupo é modelado utilizando uma densidade de probabilidade semi-paramétrico. Recentes avanços no k-means e outros algoritmos de clusterização baseados no erro médio quadrático com suas aplicações foram propostas por HANSEN & MLADENOVIAE (2001) com o **J-means**, uma nova heurística de busca local para a soma mínimos quadrados. MEILA & SHI (2001) apresentaram uma visão de clusterização espectral baseada em passeio aleatório de Markov e propuseram o **Modified Normalized Cut (MNCut),** algoritmo que pode lidar com um número arbitrário de clusters.

Em 2002, um algoritmo hierárquico incremental inovador com base na teoria da gravidade em física é apresentado por CHIEN-YU et al. (2002), o **GRIN,** um algoritmo que proporciona qualidades de clusterização favoráveis porque as configurações de parâmetros ótimos no algoritmo GRIN não são sensíveis à distribuição do conjunto de dados. BELKIN & NIYOGI (2002) apresentaram o **Laplacian Eigenmap**, outro método de clusterização espectral que deriva a representação de dados com base nos vectores próprios do grafo Laplaciano. E **ORCLUS** - arbitrarily ORiented projected CLUSter generation foi introduzido por AGGARWAL & YU (2002) e define um conjunto projetado generalizado como um subconjunto densamente distribuído de objetos de dados em um subespaço, junto com um subconjunto de vetores que representam o subespaço. BARALDI & ALPAYDIN (2002) propuseram o **SART- simplificado ART** que é descrito através de uma arquitetura feedforward combinada com um mecanismo de comparação de jogo. Como exemplos concretos, eles também propõem o ART difuso simétrico **(SFART)** e SART com redes totalmente autoorganizáveis **(FOSART).** LI et al. (2002) apresentaram o **CLINDEX**, clusterização baseada em grade para esquema de indexação, para pesquisas de semelhança aproximada em espaços de alta dimensionalidade.

O algoritmo SNN foi reformulado e misturado a uma abordagem baseada em

densidade pra dar origem ao **ROCK**, por ERTÖZ et al. (2003). Esse "novo - SNN" esparsifica a matriz de similaridade para manter os vizinhos mais próximos e deriva a força total de ligações para cada elemento. Em 2003, também foi desenvolvida abordagem bayesiana para melhorar os modelos de mistura de agregação de dados no algoritmo, **LDA- Latente Dirichlet Alocation**, por BLEI et al. (2003). CHEUNG (2003) apresentou uma outra generalização do algoritmo k-means, **o k*-means** aplicável a clusters com formatos de elipse, bem como aqueles em forma de esfera, e que realiza a clusterização sem pré determinar o número exato de clusters. LIKAS et al. (2003) propuseram um algoritmo **K-means global**, constituída por uma série de processos de clusterização k-means com o número de clusters que variam de 1 a K, independente das partições iniciais e fornece aceleração computacional. Uma extensão baseada em kernel, chamada **MSVC**–**Multi-Sphere Support Vector Clustering**, foi proposta por CHIANG & HAO (2003), que combinam o conceito de pertinência fuzzy com clusterização de suporte vetor.

BANERJEE et al. (2004) exploram a família de **distâncias de Bregman para K-means** e HARPELED & MAZUMDAR (2004) propõem um método chamado **coreset K-means,** que propõe primeiro resumir um grande conjunto em um subconjunto relativamente pequeno de dados, e em seguida, aplicar os algoritmos de clusterização ao resumidos conjunto de dados. WANG (2004) lançou o **STR- DBSCAN** (streaming DBSCAN), que é apropriado para clusterização de dados de streaming e para modelos de detecção do quadro. O **I-DBSCAN**-Improved DBSCAN proposto por BORAH & BHATTACHARYYA (2004) tem um tempo de execução melhor do que DBSCAN, mas produz mais ruído e não tabalha bem na presença de clusters de densidade variadas.

Em 2005, outro método probabilístico foi lançado um modelo probabilístico desenvolvido para agregação de dados que modela a função de densidade por um modelo probabilístico de misturas, a o **UGM-** Undirected Graphical Model, por OSINDERO et al. (2005). O **K- medoids** foi proposto em KAUFFMAN & ROUSSEEUW (2005), onde os clusters são representados usando a mediana dos dados em vez da média. Além disso, o método co-clusterização foi estendido para **clusterização multi-way** em BEKKERMAN et al. (2005) para agrupar um conjunto de objetos, agrupando simultaneamente as suas componentes heterogêneas. MOTA &

GOMIDE (2005) aplicaram o método fuzzy c-Means na fase de avaliação do cromossomo em Algoritmo Genético, lançando o método **C-Means com GA**. CAMASTRA & VERRI (2005) apresentam **NKM**- Novel Kernel Method de Clusterização baseado em Kernel que usa SVM-Support Vector Machine e inspiração no K-means para obter superfícies de separação naturalmente não-lineares dos dados.

Outra abordagem bayesiana para melhorar os modelos de mistura de agregação de dados foi desenvolvida em 2006 no modelo de **Alocação de Pachinko**, por LI & MCCALLUM (2006). Nesse ano também duas versões do DBSCAN foram lançadas para melhorar o tempo de execução ou o tratamento de dados com diferentes densidades, usando um sistema de indexação: um método de aglomeração híbrido baseado em densidade e amostragem rápido, o **l-DBSCAN**, proposto por VISWANATH & PINKESH (2006) e o **Fast-DBSCAN** introduzido por LIU (2006), baseado em densidade e kernel. Em YIN et al. (2006), os autores propuseram o **EBABS** – Encoded Bitmap Approach Based Swap, para melhorar o método hierárquico clássico, ideal para usar método de clusterização para analisar séries temporais .

O ano de 2007 foi importante para inovações em DBSCAN. Neste ano foram lançados cinco versões: o **LSH- DBSCAN** por ZHANG et al. (2007) ; o **VDBSCAN - Variado DBSCAN** por LIU et al. (2007), o **CDBSCAN -** DBSCAN com restrições por RUIZ et al. (2007); o **STDBSCAN -** DBSCAN para Dados Espaciais Temporais, por BIRANT & CUT (2007) e o **PrPrDBSCAN** - que preserva a privacidade DBSCAN, por LIU et al. (2007) . Os métodos k-means e DENCLUE também ganharam inovações neste ano: ARTHUR & VASSILVITSKII (2007) lançaram o **k-means++**, e **Denclue 2.0** foi lançado por HINNEBURG & GABRIEL (2007). OLIVEIRA (2007) propôs o uso do algoritmo EDA em um algoritmo evolutivo para Análise de Clusters com base na densidade e grade, lançando o **EDACluster**. Além disso, BANERJEE et al. (2007) apresentaram o método de clusterização baseada em grafos **MWRG-Multi-Way in Relation Graphs** para dados relacionais, onde diferentes tipos de entidades são simultaneamente agrupados com base em seus valores de atributos intrínsecos. LIU et al. (2007) introduziram o **Boostcluster,** clusterização impulsionada por restrições de pares, que melhoraram de forma iterativa a precisão de qualquer algoritmo de clusterização explorando as restrições de pares .

Em 2008 foram apresentadas mais inovações em DBSCAN. **Fast Parzem - Window DBSCAN**, por BABU & VISWANATH (2008), uma abordagem de clusterização híbrida baseado em densidade e kernel; e o **NUDBSCAN -** Non Uniform DBSCAN, por SANG & YI (2008), uma técnica que tenta superar o problema da DBSCAN suavizando a diferença de densidade entre os clusters Também em 2008 foi introduzido o **DDSC** - Density Differenciated Spatial Clustering, por BORAH & BHATTACHARYYA (2008). Na área de Redes Neurais, em 2008 foi lançado **o Método de Clusterização Baseada em Redes Complexas e Computação Bioinspirada** em OLIVEIRA (2008), um método de clusterização com base na identificação de comunidades em redes complexas e modelos computacionais biologicamente inspirados. Também em 2008, uma variante do **K-means utilizando L1 distância** foi proposta pelo KASHIMA et al. (2008).

Em 2009, MAHRAN & MAHAR (2009), apresentaram o **DBSCAN-Grid**, em que o desempenho de DBSCAN é aumentado usando grade para particionar o conjunto de dados e, em seguida, fundir os resultados.

Em HE & PAN (2010), os autores introduziram a clusterização **DENCLUE NEURO-FUZZY**, uma abordagem fuzzy para o DENCLUE . O método mostrou ser eficiente, mas não é recomendado para dados com ruídos. FERREIRA et al. (2010) propuseram o método de Clusterização **GA-Hibrido**, um método baseado em Grade, Densidade e algoritmos genéticos. LI & WANG (2010) introduziram o Algoritmo de Clusterização **SCAHIPAT** – um método particional combinado com árvores hierárquicas. PASCUCCI et al. (2010) apresentaram o primeiro método de clusterização baseado em Topologia **TODA** – Topological Data Analysis e neste ano foi lançada mais uma inovação para DBSCAN, o **DVBSCAN** por RAM et al. (2010). Além desses, um método inovador foi apresentado por XAVIER (2010) chamada **Suavização Hiperbólica**, que considera uma solução de otimização para o problema original do mínimo da soma dos quadrados da clusterização.

Em 2011 foram documentados três métodos: o **DBSCAN - Hamming** por MIMAROGLU & AKSEHIRLI (2011), em que os autores aplicam a distância de Hamming no DBSCAN; o **HyCARCE** - Hyperellipsoidal Clustering Algorithm para Resource- Constrained Ambients por MOSHTAGHI et al. (2011), que mostrou

performance superior ao DENCLUE, e o **SHMSVC- Statistical Histogram Based Multi-sphere Support Vector Clustering**, por JIA et al. (2011), um método baseado em histograma combinado ao método MSVC, adequado para grandes conjuntos de dados, robusto a ruídos e outliers e capaz de determinar automaticamente o número de clusters e identificar pontos de dados com contornos arbitrários de contornos do cluster.

Em 2012, um método de clusterização robusto contra mudanças frequentes na topologia da rede, chamado **SBDC** Schelling Based Distributed Clustering, foi introduzido. O método é inspirado no modelo de segregação de Schelling, popular na sociologia e foi proposto por TSUGAWA et al. (2012).

Em 2013, a versão mais recente do DBSCAN, **DBSCAN Revisado**, foi lançado em TRAN et al. (2013). O método tem um desempenho robusto para conjuntos de dados contendo estruturas densas com grupos conectados. Neste método, os resultados da Clusterização não dependem da ordem em que os objetos são processados. Esta é uma versão usada nesta tese.

O Quadro 1 a seguir traz a evolução dos métodos de clusterização citados, ordenados pelo ano de publicação.

Quadro 4.1: Evolução dos Métodos de Clusterização.

Ano	Método de Clusterização	Característica de Formação	Informações	Publicado em
1948	Hierárquico Ligação Completa	Hierárquico	Formação Aglomerativa	SORENSEN (1948)
1956	K-means	Particional	Realocação iterativa baseada em erro quadrático	STEINHAUS (1956)
1957	Hierárquico Ligação Simples	Hierárquico	Formação Aglomerativa	SNEATH (1957)
1963	WARD	Hierárquico	Minimiza o total de variância intra-clusters	WARD (1963)
1965	FORGY	Particional	Variante do k-means	FORGY (1965)
1965	ISODATA	Particional	Variante do K-means	BALL & HALL (1965)
1972	Hartigan	Co-clusterização	Clusterização Bidimensional	HARTINGAN (1972)
1972	Nuvens Dinâmicas	Não Hierárquico	Define as formas fortes e tipologias de um	DIDAY (1972)

			conjunto de curvas.	
1973	SNN	Baseado em Compartilhamento do Vizinho mais Próximo	Boa performance em ruído	JARVIS & PATRICK (1973)
1973	Fuzzy C-means	Fuzzy particional		DUNN (1973)
1978	BACKER	Fuzzy particional		BACKER (1978)
1980	K-means usando distância Itakura-Saito	Particional		LINDE et al. (1980)
1981	Fuzzy C-means melhorado	Fuzzy particional		BEZDEK (1981)
1983	S.A	Baseado em Pesquisa Combinatoria	Particional Simulated Annealling	KIRKPATRICK et al. (1983)
1987	COBWEB	Hierárquico, Baseado em Modelos	Organiza os dados em uma árvore de forma incremental	FISHER (1987)
1988	JAIN	Baseado em Densidade		JAIN (1988)

Ano	Método de Clusterização	Característica de Formação	Informações	Publicado em
1988	ART	Baseado em Redes Neurais	Teoria ressonante adaptativa	CARPENTER et al. (1988)
1989	SOM- Mapa de Kohonnen	Baseado em Redes Neurais	Mapas auto organizáveis	KOHONEN (1989)
1989	TS cluster	Baseado em Pesquisa Combinatoria	Tabu Search	GLOVER (1989)
1989	PAM	Particional	Particional em torno de medóides, com realocacao iterativa	KAUFFMAN & ROUSSEEUW (1989)
1989	CLARA		PAM melhorado para grandes volumes de dados	
1989	DIANA	Hierárquico	Formação Divisiva	
1989	MONA		Análise Monotética	
1989	AGNES		AGglomerative NESting	
1989	CLASSIT	Hierárquico Incremental Baseado em Modelos	Extensão do Método Cobweb para dados contínuos	FISHER et al. (1989)
1992	RatioCut	Baseado em Teoria de Grafos		HAGEN & KAHNG (1992)
1993	FCS	Fuzzy particional	Fuzzy C-Sheels	YANG (1993)
1993	PCM	Fuzzy Probabilístico	Probabilístico C-Means	KRISHNAPURAM & KELLER (1993)
1994	CLARANS	Baseada em Pesquisa Aleatória	Clustering Large Applications based on RANdomized Search)	HAN & NG (1994)
1994	Clusterização de Montanha	Fuzzy	Estima os centros dos clusters	YAGER & FILEV (1994)
1994	EJCluster	Hierárquico	Formação Divisiva	GARCÍA et al. (1994)
1995	OPT K-means	Particional	Variante do K-means	CHINRUNGRUG & SÉQUIN (1995)
1995	R-TREE	Clusterização Otimizada		ESTER et al. (1995)
1996	DBSCAN	Baseado em Densidade	Dedicado para dados com ruído	ESTER et al. (1996)
1996	K-means com distância Mahalanobis	Particional	Variante do K-means	MAO & JAIN (1996)
1996	HEC	Baseado em Redes Neurais	Gaussian Mixture Densities Decomposition	

Ano	Método de Clusterização	Característica de Formação	Informações	Publicado em
1996	Gridfile	Hierárquico Baseado em Grades	Para base de dados muito grande	SCHIKUTA (1996)
1996	BIRCH	Hierárquico	Formação Aglomerativa Otimizada	ZHUANG et al. (1996)
1996	GMDD	Particional Baseado em Modelo Probabilístico	Gaussian Mixture Densities Decomposition	ZHUANG et al. (1996)
1996	AUTOCLASS	Particional Baseado em Modelos Probabilístico	Usa estatística bayesiana para estimar o numero de clusters	CHEESEMAN & STUTZ (1996)
1997	GDBSCAN	Baseado em Densidade	DBSCAN Generalizado	ESTER et al. (1997)
1997	STING	Hierárquico aglomerativo e baseado em grades	STatistical INformation Grid Clusterização Otimizada	WANG et al. (1997)
1997	EM -expectation maximization	Particional Probabilístico	Generalização do k-means e do fuzzy c-means	MCLACHLAN et al. (1997)
1997	HMM	Clusterização de Sequências	Hidden Markov Model	MOZER et al. (1997
1998	k-modes	Particional	Usa a medida do coediciente de semelhança	HUANG (1998)
1998	k-prothotypes		Integra k-means e k-modes	
1998	DENCLUE	Baseado em Densidade	Detecta clusters de formato arbirtrário	HINNEBURG & KEIM (1998)
1998	CLIQUE	Baseado em Densidade e Grade	Bom para dados com grandes dimensões	AGRAWAL et al. (1998)
1998	DBCLASD	Baseado em Grade e em Modelos	Usa a base teórica do teste qui quadrado	XU et al. (1998)
1998	WAVE CLUSTER	Baseado em wavelet, grade e densidade		SHEIKHOLESLAMI et al. (1998)
1998	K-MEANS kernel	Particional Baseado em Kernel	Detecta clusters de formato arbitrário	SHOLKÖPF et al. (1988)
1998	CURE	Hierárquico	Formação Aglomerativa	GUHA et al. (1998)

Ano	Método de Clusterização	Característica de Formação	Informações	Publicado em
1998	DBSCAN Incremental	Baseado em Densidade	Versão Incremental do DBSCAN	ESTER et al. (1998)
1998	Fast K-means	Particional	Versão incremental, rápida e escalável de uma única passagem de K-means	BRADLEY et al. (1998)
1999	GGA	Baseado em Computação Evolucionária	Genetical Guided Algorithm	HALL et al. (1999)
1999	GKA	Baseado em Computação Evolucionária	Genetic k-means Algorithm	KRISHNA & MURTYG (1999)
1999	OPTICS	Baseado em Densidade	Ordering Points To Identify the Clustering Structure	ANKERST et al. (1999)
1999	CHAMALEON	Hierárquico Baseado em Grafos	Formação Aglomerativa	KARYPIS et al. (1999)
1999	FDC	Baseado em Densidade	Fast Density Clustering	BO ZHOU et al. (1999)
1999	PDBSCAN	Baseado em Densidade	Paralel DBSCAN	XU et al. (1999).
1999	MAFIA	Hierárquico	Merging Adaptative Finite Intervals	HARASHA et al. (1999)
1999	K-means kd-tree	Particional	Variante do K-means	PELLEG E MOORE (1999)
2000	X-Means	Particional	Variante do K-means	PEELEG & MOORE (2000)
2000	K-Means Divisivo	Hierárquico Particional	Abordagem Divisiva	STEINBACH et al. (2000)
2000	HCS	Baseado em Grafos	Highly Connected Subgraphs	HARTUV et al. (2000)
2000	ROCK	Hierárquico Aglomerativo	Clusterização Robusta Usando Ligações	GUHA et al. (2000)
2000	Normalized Cut	Baseado em Modelo		SHI & MALIK (2000)
2000	Divide and conquer	Hierárquico	Sumarização de dados	STEINBACH et al. (2000)
2000	SB-DBSCAN	Baseado em Densidade	Variante do DBSCAN	ZHOU et al. (2000)
2000	S-DBSCAN	Baseado em Densidade e Amostragem		
2001	COD	Baseado em Restrições	Clusterizacao com distancias obstruidas	TUNG et al. (2001)
Ano	**Método de**	**Característica de**	**Informações**	**Publicado em**

	Clusterização	Formação		
2001	Método da Entropia Mínima	Baseado em Modelos		ROBERTS et al. (2001)
2001	SVC	Particional baseado em Kernel	Clusterizacao de suporte vetor	BEM HUR et al. (2001)
2001	J-means	Particional	Variante do k-means	HANSEN & MLADENOVIAE (2001)
2001	MNCut	Clusterizaçãoi Espectral	Modified Normalized Cut	MEILA & SHI (2001)
2002	GRIN	Hierárquico Incremental	Baseado na Teoria da Gravidade	CHIEN-YU et al. (2002)
2002	Laplacian Eingenmap	Clusterização Espectral	Baseado nos autovetores do grafo laplaciano	BELKIN & NIYOGI 2002)
2002	ORCLUS	Baseado em Grafos	arbitrarily ORiented projected CLUSter generation	AGGARWAL & YU (2002)
2002	SART	Baseado em Redes Neurais	Simplificado ART	BARALDI & ALPAYDIN (2002)
2002	SFART		ART Simétrico Difuso	
2002	FOSART		SART com redes totalmente autoorganizáveis	
2002	CLINDEX	Baseado em Grades	Clusterização para para esquema de indexação	LI et al. (2002)
2003	K*-means	Particional	Variante do k-means	CHEUNG (2003)
2003	K-Means Global			LIKAS et al. (2003)
2003	ROCK	Baseado em Compartilhamento do Vizinho mais Próximo	Novo SNN	ERTÖZ et al. (2003)
2003	LDA	Baseado em Estatística Bayesiana	Latent Dirichlet Allocation	BLEI et al. (2003)
2003	MSVC	Baseado em lógica fuzzy e clusterização de suporte vetor	Multi-Sphere Support Vector Clustering	CHIANG & HAO (2003)
2004	K-means com distâncias de Bregman	Particional	Variante do k-means	BANERJEE et al. (2004)

Ano	Método de Clusterização	Característica de Formação	Informações	Publicado em
2004	CORESET k-means	Particional Baseado em Resumo de Dados	Variante do k-means	HAR-PELED & MAZUMDAR (2004)
2004	STR- DBSCAN	Baseado em Densidade	Streaming DBSCAN	WANG (2004)
2004	I-DBSCAN		Improved DBSCAN	BORAH & BHATTACHARYYA (2004)
2005	MULTI WAY CLUSTERING	Extensão do co-Clusterização		BEKKERMAN et al. (2005)
2005	K-medoids	Particional	Clusters representados pela mediana	KAUFMAN & ROUSSEEU (2005)
2005	UGM	Baseado em Modelos	Undirected Graphical model	OSINDERO et al. (2005).
2005	C-Means com GA	Baseado em lógica fuzzy e algoritmos genéticos		MOTA & GOMIDE (2005)
2005	NKM	Baseado em Kernel e clusterização de suporte vetor	Novel Kernel Method	CAMASTRA & VERRI (2005)
2006	Pachinko Allocation Model	Baseado em Estatística Bayesiana	Melhora os modelos de mistura de agregação de dados	LI & MCCALLUM (2006)
2006	l-DBSCAN	Baseado em amostragem e densidade	Variante do DBSCAN	VISWANATH & PINKESH (2006)
2006	Fast DBSCAN	Baseado em Kernel e Densidade	Variante do DBSCAN	LIU (2006)
2006	EBABS	Hierárquico Aglomerativo	ENCODED BITMAP APPROACH BASED	YIN et al. (2006)
2007	ST-DBSCAN	Baseado em Densidade	DBSCAN para Dados Temporais	BIRANT & CUT (2007)
2007	VDBSCAN		DBSCAN Variado	LIU et al. (2007)
2007	C-DBSCAN		DBSCAN para restrições	RUIZ et al. (2007)
2007	LSH DBSCAN		DIMINUI O TEMPO DE EXECUÇÃO DO DBSCAN	ZHANG et al. (2007)
2007	PrPr-DBSCAN		Preserva Privacidade DBSCAN	LIU et al. (2007)
2007	DENCLUE 2.0		Denclue Melhorado	HINNEBURG E GABRIEL (2007)

Ano	Método de Clusterização	Característica de Formação	Informações	Publicado em
2007	k-means ++	Particional	Variante do K-means	ARTHUR & VASSILVITSKII (2007)
2007	EDACluster	Baseado em Densidade e Grade	Algoritmo evolutivo EDA	OLIVEIRA (2007)
2007	MWRD-Multi-Way	Baseada em Grafos	Multi Way in Relation Graphs	BANERJEE et al. (2007)
2007	Boostcluster	Baseado em Restrições	Boosting framework Para Clusterização de Dados	LIU et al. (2007)
2008	FPW-DBSCAN	Baseado em Kernel e Densidade	Fast Parzem -Window DBSCAN	BABU & VISWANATH (2008)
2008	NUDBSCAN	Baseado em Densidade	Non Uniform DBSCAN	SANG & YI (2008)
2008	DDSC	Baseado em Densidade	Density Differenciated Spatial Clustering	BORAH & BHATTACHARYYA (2008)
2008	Redes complexas e computação Bioinspirada	Baseada em Computação Evolucionária		OLIVEIRA (2008)
2009	Grid- DBSCAN	Baseado em Densidade e Grade	Variante do DBSCAN	MAHRAN & MAHAR (2009)
2010	DVBSCAN	Baseado em Densidade	Variante do DBSCAN	RAM et al. (2010)
2010	DENCLUE NEURO FUZZY	Baseado em Densidade, redes neurais e lógicafuzzy	Eficiente mas não adequado para dados com ruído	HE & PAN (2010)
2010	GA HIBRIDO	Baseado em Densidade, em Grade e Algoritmos Genéticos		FERREIRA et al. (2010)
2010	SCAHIPAT	Hierarquico Particional	Spatial Clusterização Algorithm Based on Hierarchical-Partition Tree	LI & WANG (2010)
2010	TODA	Baseado em Topologia	Topological Data Analysis	PASCUCCI et al. (2010)
2010	SUAVIZAÇÃO HIPERBÓLICA	Baseado em Otimização		XAVIER (2010)

Ano	Método de Clusterização	Característica de Formação	Informações	Publicado em
2011	DBSCAN-Hamming	Baseado em Densidade	DBSCAN usando a distancia de Hamming	MIMAROGLU & AKSEHIRLI (2011)
2011	HyCARCE	Baseado em Restrições	Hyperellipsoidal Clustering Algorithm for Resource-Constrained Environments	MOSHTAGHI et al. (2011)
2011	SHMSVC	Baseado em modelo e clusterização de suporte vetor	Statistical Histogram Based Multi-sphere Support Vector Clustering,	JIA et al. (2011)
2012	SBDC	Baseado em Conectividade	Schelling Based Distributed Clustering	TSUGAWA et al. (2012).
2013	DBSCAN Revisado	Baseado em Densidade	Resolve o problema de detecção de objetos de fronteira de cluster	TRAN et al. (2013)