# Análise de cluster

Fonte: <https://stringfixer.com/pt/Cluster_Analysis>

**A análise de cluster** ou **clustering** é a tarefa de agrupar um conjunto de objetos de forma que os objetos no mesmo grupo (chamado de **cluster** ) sejam mais semelhantes (em certo sentido) entre si do que em outros grupos (clusters). É uma tarefa principal de [análise exploratória de dados](https://stringfixer.com/pt/Exploratory_data_analysis) e uma técnica comum para [análise](https://stringfixer.com/pt/Data_analysis) [estatística de](https://stringfixer.com/pt/Statistics) [dados](https://stringfixer.com/pt/Data_analysis) , usada em muitos campos, incluindo [reconhecimento de padrões](https://stringfixer.com/pt/Pattern_recognition) , [análise de imagens](https://stringfixer.com/pt/Image_analysis) , [recuperação de informações](https://stringfixer.com/pt/Information_retrieval) , [bioinformática](https://stringfixer.com/pt/Bioinformatics) , [compressão de dados](https://stringfixer.com/pt/Data_compression) , [computação gráfica](https://stringfixer.com/pt/Computer_graphics) e [aprendizado de máquina](https://stringfixer.com/pt/Machine_learning) .

Uma imagem contendo relógio, objeto, lendo, escuro

Descrição gerada automaticamente

O resultado de uma análise de agrupamento mostrado como a coloração dos quadrados em três agrupamentos.

A análise de cluster em si não é um [algoritmo](https://stringfixer.com/pt/Algorithm) específico , mas a tarefa geral a ser resolvida. Isso pode ser alcançado por vários algoritmos que diferem significativamente em sua compreensão do que constitui um cluster e como encontrá-los de forma eficiente. Noções populares de clusters incluem grupos com pequenas [distâncias](https://stringfixer.com/pt/Distance_function) entre os membros do cluster, áreas densas do espaço de dados, intervalos ou [distribuições estatísticas](https://stringfixer.com/pt/Statistical_distribution) particulares . O agrupamento pode, portanto, ser formulado como um problema de [otimização multi-objetivo](https://stringfixer.com/pt/Multi-objective_optimization) . O algoritmo de agrupamento apropriado e configurações de parâmetro (incluindo parâmetros como a [função de distância](https://stringfixer.com/pt/Metric_(mathematics)) a ser usada, um limite de densidade ou o número de clusters esperados) dependem do [conjunto de dados](https://stringfixer.com/pt/Data_set) individual e do uso pretendido dos resultados. A análise de cluster como tal não é uma tarefa automática, mas um processo iterativo de [descoberta](https://stringfixer.com/pt/Knowledge_discovery) de [conhecimento](https://stringfixer.com/pt/Knowledge_discovery) ou otimização multi-objetivo interativa que envolve tentativa e falha. Freqüentemente, é necessário modificar o pré-processamento de dados e os parâmetros do modelo até que o resultado atinja as propriedades desejadas.

Além do termo *agrupamento* , há vários termos com significados semelhantes, incluindo [*classificação*](https://stringfixer.com/pt/Statistical_classification) *automática* , [*taxonomia numérica*](https://stringfixer.com/pt/Numerical_taxonomy) , *botriologia* (do grego βότρυς "uva"), *análise tipológica* e [*detecção de comunidade*](https://stringfixer.com/pt/Community_structure) . As diferenças sutis estão frequentemente no uso dos resultados: enquanto na mineração de dados, os grupos resultantes são o assunto de interesse, na classificação automática o poder discriminativo resultante é de interesse.

A análise de agrupamento foi originada na antropologia por Driver e Kroeber em 1932 [[1]](https://stringfixer.com/pt/Cluster_Analysis) e introduzida na psicologia por [Joseph Zubin](https://stringfixer.com/pt/Joseph_Zubin) em 1938 [[2]](https://stringfixer.com/pt/Cluster_Analysis) e [Robert Tryon](https://stringfixer.com/pt/Robert_Tryon) em 1939 [[3]](https://stringfixer.com/pt/Cluster_Analysis) e famosa por [Cattell a](https://stringfixer.com/pt/Raymond_Cattell) partir de 1943 [[4]](https://stringfixer.com/pt/Cluster_Analysis) para classificação da teoria dos traços em [psicologia da personalidade](https://stringfixer.com/pt/Personality_psychology) .

## Definição

A noção de um "cluster" não pode ser definida com precisão, o que é uma das razões pelas quais existem tantos algoritmos de clustering. [5] Há um denominador comum: um grupo de objetos de dados. No entanto, diferentes pesquisadores empregam diferentes modelos de cluster, e para cada um desses modelos de cluster, novamente, diferentes algoritmos podem ser fornecidos. A noção de cluster, conforme encontrada por diferentes algoritmos, varia significativamente em suas propriedades. Compreender esses "modelos de cluster" é a chave para entender as diferenças entre os vários algoritmos. Os modelos de cluster típicos incluem:

* *Conectividade modelo s*: por exemplo,agrupamento hierárquicoconstrói modelos baseados em conectividade distância.
* *Modelo centróide s*: por exemplo, oalgoritmo k-médiasrepresenta cada cluster por um único vetor médio.
* *Modelos de distribuição s*: os clusters são modelados usando distribuições estatísticas, comodistribuições normais multivariadasusadas peloalgoritmo de maximização de expectativa.
* *Densidade modelo s*: por exemplo,DBSCANeÓPTICAdefine aglomerados como regiões densas ligados no espaço de dados.
* *Subespaço modelo s*: embiclustering(também conhecido como co-agregação ou de dois modos de agrupamento), aglomerados são modelados com ambos os membros de fragmentação e atributos relevantes.
* *Grupo modelo s*: alguns algoritmos não fornecem um modelo refinado de seus resultados e fornecer apenas as informações de agrupamento.
* *Baseado no gráfico modelo s*: umgrupo exclusivo, que é, um subconjunto de nós de umgráficode tal modo que cada dois nós no subconjunto estão ligados por uma aresta pode ser considerado como uma forma de protótipo do conjunto. Relaxamentos do requisito de conectividade completo (uma fração das arestas pode estar faltando) são conhecidos como quase-cliques, como noalgoritmo de agrupamento HCS.
* *Modelos de gráfico assinado* : cada caminho em um gráfico assinado tem um sinal do produto dos sinais nas bordas. De acordo com os pressupostos da teoria do equilíbrio , as arestas podem mudar de sinal e resultar em um gráfico bifurcado. O "axioma de capacidade de agrupamento" mais fraco (nenhum ciclo tem exatamente uma borda negativa) produz resultados com mais de dois clusters, ou subgráficos com apenas bordas positivas. [6]
* *Neural modelo s*: o mais conhecidosem supervisão rede neuralé omapa de auto-organizaçãoe esses modelos podem geralmente ser caracterizado como semelhante a um ou mais dos modelos acima, e incluindo modelos subespaciais quando as redes neurais implementar uma forma deAnálise de Componentes PrincipaisouAnálise de componentes independentes.

Um "clustering" é essencialmente um conjunto de tais clusters, geralmente contendo todos os objetos no conjunto de dados. Além disso, pode especificar o relacionamento dos clusters entre si, por exemplo, uma hierarquia de clusters embutidos uns nos outros. Os agrupamentos podem ser distinguidos aproximadamente como:

* *Clustering rígido* : cada objeto pertence a um cluster ou não
* *Clustering suave* (também:*agrupamento difuso* ): cada objeto pertence a cada agrupamento em um determinado grau (por exemplo, uma probabilidade de pertencer ao agrupamento)

Também existem distinções mais sutis possíveis, por exemplo:

* *Clustering de particionamento estrito* : cada objeto pertence a exatamente um cluster
* *Clustering de particionamento estrito com outliers* : os objetos também não podem pertencer a nenhum cluster e são consideradosoutliers
* *Clustering sobreposto* (também:*clustering alternativo*,*clusteringmulti-view*): os objetos podem pertencer a mais de um cluster; geralmente envolvendo clusters rígidos
* *Clustering hierárquico* : objetos que pertencem a um cluster filho também pertencem ao cluster pai
* *Clustering de subespaço* : enquanto um clustering de sobreposição, dentro de um subespaço definido exclusivamente, não se espera que os clusters se sobreponham

## Algoritmos

Conforme listado acima, os algoritmos de cluster podem ser categorizados com base em seu modelo de cluster. A visão geral a seguir listará apenas os exemplos mais proeminentes de algoritmos de agrupamento, pois há possivelmente mais de 100 algoritmos de agrupamento publicados. Nem todos fornecem modelos para seus clusters e, portanto, não podem ser facilmente categorizados. Uma visão geral dos algoritmos explicados na Wikipedia pode ser encontrada na lista de algoritmos estatísticos .

Não existe um algoritmo de agrupamento objetivamente "correto", mas, como foi observado, "o agrupamento está nos olhos de quem vê". [5] O algoritmo de agrupamento mais apropriado para um problema particular geralmente precisa ser escolhido experimentalmente, a menos que haja uma razão matemática para preferir um modelo de agrupamento a outro. Um algoritmo projetado para um tipo de modelo geralmente falhará em um conjunto de dados que contém um tipo de modelo radicalmente diferente. [5] Por exemplo, k-means não pode encontrar clusters não convexos. [5]

## Clustering baseado em conectividade (clustering hierárquico)

O clustering baseado em conectividade, também conhecido como clustering hierárquico , é baseado na ideia central de que os objetos estão mais relacionados a objetos próximos do que a objetos mais distantes. Esses algoritmos conectam "objetos" para formar "clusters" com base em sua distância. Um cluster pode ser descrito amplamente pela distância máxima necessária para conectar partes do cluster. Em diferentes distâncias, diferentes clusters se formarão, que podem ser representados usando um dendrograma , que explica de onde vem o nome comum " agrupamento hierárquico ": esses algoritmos não fornecem um único particionamento do conjunto de dados, mas em vez disso fornecem uma extensa hierarquia de aglomerados que se fundem em certas distâncias. Em um dendrograma, o eixo y marca a distância na qual os clusters se fundem, enquanto os objetos são colocados ao longo do eixo x de forma que os clusters não se misturem.

O clustering baseado em conectividade é uma família inteira de métodos que diferem pela maneira como as distâncias são calculadas. Além da escolha usual de funções de distância , o usuário também precisa decidir sobre o critério de ligação (uma vez que um cluster consiste em vários objetos, existem vários candidatos para calcular a distância) a usar. As escolhas populares são conhecidas como agrupamento de ligação única (o mínimo de distâncias de objetos), agrupamento de ligação completa (o máximo de distâncias de objetos) e UPGMA ou WPGMA ("Método de grupo de pares não ponderados ou ponderados com média aritmética", também conhecido como ligação média clustering). Além disso, o clustering hierárquico pode ser aglomerativo (começando com elementos únicos e agregando-os em clusters) ou divisivo (começando com o conjunto de dados completo e dividindo-o em partições).

Esses métodos não produzirão um particionamento exclusivo do conjunto de dados, mas uma hierarquia da qual o usuário ainda precisa escolher os clusters apropriados. Eles não são muito robustos em relação a outliers, que aparecerão como clusters adicionais ou até mesmo farão com que outros clusters se fundam (conhecido como "fenômeno de encadeamento", em particular com clustering de ligação única ). No caso geral, a complexidade é para agrupamento aglomerativo e para clustering divisivo , [7] o que os torna muito lentos para grandes conjuntos de dados. Para alguns casos especiais, métodos eficientes ideais (de complexidade) são conhecidos: SLINK [8] para ligação única e CLINK [9] para agrupamento de ligação completa. Na comunidade de mineração de dados , esses métodos são reconhecidos como uma base teórica da análise de agrupamento, mas frequentemente considerados obsoletos [ carece de fontes? ] . No entanto, eles forneceram inspiração para muitos métodos posteriores, como agrupamento baseado em densidade.

Exemplos de clustering de ligação

Luz do sol

Descrição gerada automaticamente com confiança baixa

Ligação única em dados Gaussianos. Com 35 clusters, o maior cluster começa a se fragmentar em partes menores, enquanto antes ainda estava conectado ao segundo maior devido ao efeito de link único.

Uma imagem contendo luz, estrela, pare, tráfego

Descrição gerada automaticamente

Ligação única em clusters baseados em densidade. 20 clusters extraídos, a maioria dos quais contendo elementos únicos, uma vez que o linkage clustering não tem noção de "ruído".

Agrupamento baseado em centróide

No agrupamento baseado em centróide, os clusters são representados por um vetor central, que pode não ser necessariamente um membro do conjunto de dados. Quando o número de clusters é fixado em *k* , o agrupamento k -means dá uma definição formal como um problema de otimização: encontre os *k* centros do cluster e atribua os objetos ao centro do cluster mais próximo, de forma que as distâncias quadradas do cluster sejam minimizadas.

O problema de otimização em si é conhecido como NP-difícil e, portanto, a abordagem comum é buscar apenas soluções aproximadas. Um método aproximado particularmente bem conhecido é o algoritmo de Lloyd , [10] frequentemente referido apenas como " *algoritmo k-means* " (embora outro algoritmo tenha introduzido esse nome ). No entanto, ele encontra apenas um ótimo local e normalmente é executado várias vezes com diferentes inicializações aleatórias. Variações de *k-* médias geralmente incluem otimizações como escolher o melhor de várias execuções, mas também restringir os centróides a membros do conjunto de dados ( k- medianas ), escolher medianas ( agrupamento de k- medianas ), escolher os centros iniciais de forma menos aleatória ( k -means ++ ) ou permitindo uma atribuição de cluster difuso ( fuzzy c-means ).

A maioria dos algoritmos do tipo *k* -means requer que o número de clusters - *k* - seja especificado com antecedência, o que é considerado uma das maiores desvantagens desses algoritmos. Além disso, os algoritmos preferem clusters de tamanho aproximadamente semelhante, pois eles sempre atribuirão um objeto ao centróide mais próximo. Isso geralmente leva ao corte incorreto das bordas dos clusters (o que não é surpreendente, pois o algoritmo otimiza os centros dos clusters, não as bordas dos clusters).

K-means tem várias propriedades teóricas interessantes. Primeiro, ele particiona o espaço de dados em uma estrutura conhecida como diagrama de Voronoi . Em segundo lugar, é conceitualmente próximo da classificação do vizinho mais próximo e, como tal, é popular no aprendizado de máquina . Terceiro, pode ser visto como uma variação do agrupamento baseado em modelo e o algoritmo de Lloyd's como uma variação do algoritmo de maximização de expectativa para este modelo discutido abaixo.

* *k -* significa exemplos de agrupamento
* Estrela dourada em fundo preto

  Descrição gerada automaticamente com confiança média

*k-* significa que separa os dados em células de Voronoi, que assume clusters de tamanhos iguais (não é adequado aqui)

* Fogos de artifício à noite

  Descrição gerada automaticamente com confiança média

*k* -means não pode representar clusters baseados em densidade

Problemas de agrupamento baseados em centróide, como *k-* médios e *k-* médios, são casos especiais do problema de localização de instalação métrica não capacitada , um problema canônico na pesquisa de operações e comunidades de geometria computacional. Em um problema de localização de instalação básica (no qual existem inúmeras variantes que modelam configurações mais elaboradas), a tarefa é encontrar os melhores locais de depósito para atender de forma otimizada um determinado conjunto de consumidores. Pode-se ver "armazéns" como centróides do cluster e "locais do consumidor" como os dados a serem agrupados. Isso torna possível aplicar as soluções algorítmicas bem desenvolvidas da literatura de localização de instalações para o problema de agrupamento baseado em centróide atualmente considerado.

Clustering baseado em distribuição

O modelo de armazenamento em cluster mais relacionado às estatísticas é baseado em modelos de distribuição . Os clusters podem então ser facilmente definidos como objetos pertencentes mais provavelmente à mesma distribuição. Uma propriedade conveniente dessa abordagem é que ela se parece muito com a maneira como os conjuntos de dados artificiais são gerados: amostrando objetos aleatórios de uma distribuição.

Embora a base teórica desses métodos seja excelente, eles sofrem de um problema-chave conhecido como overfitting , a menos que sejam impostas restrições à complexidade do modelo. Um modelo mais complexo geralmente será capaz de explicar os dados melhor, o que torna a escolha da complexidade do modelo apropriada inerentemente difícil.

Um método proeminente é conhecido como modelos de mistura gaussiana (usando o algoritmo de maximização de expectativa ). Aqui, o conjunto de dados é geralmente modelado com um número fixo (para evitar overfitting) de distribuições gaussianas que são inicializadas aleatoriamente e cujos parâmetros são otimizados iterativamente para melhor se ajustar ao conjunto de dados. Isso irá convergir para um ótimo local , portanto, várias execuções podem produzir resultados diferentes. Para obter um agrupamento rígido, os objetos são frequentemente atribuídos à distribuição gaussiana à qual provavelmente pertencem; para agrupamentos suaves, isso não é necessário.

O armazenamento em cluster baseado em distribuição produz modelos complexos para clusters que podem capturar correlação e dependência entre atributos. No entanto, esses algoritmos colocam uma carga extra no usuário: para muitos conjuntos de dados reais, pode não haver um modelo matemático definido de forma concisa (por exemplo, assumir distribuições gaussianas é uma suposição bastante forte sobre os dados).

* Exemplos de agrupamento de modelos de mistura gaussiana
* Luz brilhando no escuro

  Descrição gerada automaticamente com confiança média

Em dados distribuídos gaussianos, EM funciona bem, uma vez que usa gaussianas para modelar clusters

* Imagem aproximada de luz

  Descrição gerada automaticamente com confiança média

Os clusters baseados em densidade não podem ser modelados usando distribuições Gaussianas

Agrupamento baseado em densidade

No agrupamento baseado em densidade, [11] os clusters são definidos como áreas de maior densidade do que o restante do conjunto de dados. Objetos em áreas esparsas - que são necessários para separar clusters - geralmente são considerados ruídos e pontos de fronteira.

O método de agrupamento baseado em densidade mais popular [12] é o DBSCAN . [13] Em contraste com muitos métodos mais novos, ele apresenta um modelo de cluster bem definido chamado "densidade-alcançabilidade". Semelhante ao agrupamento baseado em ligação, é baseado em pontos de conexão dentro de certos limites de distância. No entanto, ele conecta apenas pontos que satisfaçam um critério de densidade, na variante original definida como um número mínimo de outros objetos dentro desse raio. Um cluster consiste em todos os objetos conectados por densidade (que podem formar um cluster de uma forma arbitrária, em contraste com muitos outros métodos) mais todos os objetos que estão dentro do alcance desses objetos. Outra propriedade interessante do DBSCAN é que sua complexidade é bastante baixa - requer um número linear de consultas de intervalo no banco de dados - e que descobrirá essencialmente os mesmos resultados (é determinístico para pontos centrais e de ruído, mas não para pontos de fronteira) em cada execução, portanto, não há necessidade de executá-lo várias vezes. OPTICS [14] é uma generalização do DBSCAN que remove a necessidade de escolher um valor apropriado para o parâmetro de intervaloe produz um resultado hierárquico relacionado ao agrupamento de links . DeLi-Clu, [15] Density-Link-Clustering combina idéias de clustering de ligação única e OPTICS, eliminando oparâmetro inteiramente e oferecendo melhorias de desempenho em relação ao OPTICS usando um índice R-tree .

A principal desvantagem do DBSCAN e do OPTICS é que eles esperam algum tipo de queda na densidade para detectar as bordas do cluster. Em conjuntos de dados com, por exemplo, distribuições gaussianas sobrepostas - um caso de uso comum em dados artificiais - as bordas do cluster produzidas por esses algoritmos muitas vezes parecerão arbitrárias, porque a densidade do cluster diminui continuamente. Em um conjunto de dados que consiste em misturas de gaussianas, esses algoritmos são quase sempre superados por métodos como agrupamento EM, que são capazes de modelar precisamente esse tipo de dados.

O deslocamento médio é uma abordagem de agrupamento em que cada objeto é movido para a área mais densa em sua vizinhança, com base na estimativa da densidade do kernel . Eventualmente, os objetos convergem para máximos locais de densidade. Semelhante ao agrupamento de k-means, esses "atratores de densidade" podem servir como representantes para o conjunto de dados, mas o deslocamento médio pode detectar clusters de formato arbitrário semelhantes ao DBSCAN. Devido ao caro procedimento iterativo e estimativa de densidade, o deslocamento médio é geralmente mais lento do que DBSCAN ou k-Means. Além disso, a aplicabilidade do algoritmo de deslocamento médio para dados multidimensionais é prejudicada pelo comportamento não uniforme da estimativa da densidade do kernel, o que resulta em sobre-fragmentação das caudas do cluster. [15]

* Exemplos de agrupamento baseado em densidade
* Imagem de vídeo game

  Descrição gerada automaticamente com confiança média

Clustering baseado em densidade com DBSCAN .

* Estrela dourada em fundo preto

  Descrição gerada automaticamente com confiança média

DBSCAN assume clusters de densidade semelhante e pode ter problemas para separar clusters próximos

* Luz brilhando no céu escuro

  Descrição gerada automaticamente com confiança baixa

OPTICS é uma variante do DBSCAN, melhorando o manuseio de clusters de diferentes densidades

Clustering baseado em grade

A técnica baseada em grade é usada para um conjunto de dados multidimensionais . [16] Nesta técnica, criamos uma estrutura de grade, e a comparação é realizada em grades (também conhecidas como células). A técnica baseada em grade é rápida e possui baixa complexidade computacional. Existem dois tipos de métodos de agrupamento baseados em grade: STING e CLIQUE. As etapas envolvidas no algoritmo de clusterização baseado em grade são:

1. Divida o espaço de dados em um número finito de células.
2. Selecione aleatoriamente uma célula 'c', onde c não deve ser percorrido de antemão.
3. Calcule a densidade de 'c'
4. Se a densidade de 'c' for maior que a densidade limite
   1. Marque a célula 'c' como um novo cluster
   2. Calcule a densidade de todos os vizinhos de 'c'
   3. Se a densidade de uma célula vizinha for maior que a densidade de limiar, então, adicione a célula no cluster e repita as etapas 4.2 e 4.3 até que não haja nenhum vizinho com uma densidade maior que a densidade de limiar.
5. Repita as etapas 2,3 e 4 até que todas as células sejam percorridas.
6. Pare.

Desenvolvimentos recentes

Nos últimos anos, um esforço considerável foi colocado para melhorar o desempenho dos algoritmos existentes. [17] [18] Entre eles estão *CLARANS* , [19] e *BIRCH* . [20] Com a necessidade recente de processar conjuntos de dados cada vez maiores (também conhecidos como big data ), a disposição de trocar o significado semântico dos clusters gerados por desempenho tem aumentado. Isso levou ao desenvolvimento de métodos de pré-clustering, como clustering canopy , que pode processar enormes conjuntos de dados de forma eficiente, mas os "clusters" resultantes são apenas um pré-particionamento bruto do conjunto de dados para então analisar as partições com métodos mais lentos existentes, como como k-means clustering .

Para dados de alta dimensão , muitos dos métodos existentes falham devido à maldição da dimensionalidade , o que torna as funções de distância específicas problemáticas em espaços de alta dimensão. Isso levou a novos algoritmos de agrupamento para dados de alta dimensão que se concentram no agrupamento de subespaço (onde apenas alguns atributos são usados ​​e os modelos de agrupamento incluem os atributos relevantes para o agrupamento) e agrupamento de correlação que também procura subespaço rotativo arbitrário ("correlacionado") clusters que podem ser modelados fornecendo uma correlação de seus atributos. [21] Exemplos para tais algoritmos de agrupamento são CLIQUE [22] e SUBCLU . [23]

Idéias de métodos de agrupamento baseados em densidade (em particular a família de algoritmos DBSCAN / OPTICS ) foram adaptados para agrupamento de subespaço (HiSC, [24] agrupamento de subespaço hierárquico e DiSH [25] ) e agrupamento de correlação (HiCO, [26] correlação hierárquica clustering, 4C [27] usando "conectividade de correlação" e ERiC [28] explorando clusters de correlação baseados em densidade hierárquica).

Vários sistemas de agrupamento diferentes baseados em informações mútuas foram propostos. Uma é a *variação da* métrica *de informação de* Marina Meilă ; [29] outro fornece clustering hierárquico. [30] Usando algoritmos genéticos, uma ampla gama de diferentes funções de ajuste pode ser otimizada, incluindo informações mútuas. [31] Também a propagação de crenças , um desenvolvimento recente na ciência da computação e física estatística , levou à criação de novos tipos de algoritmos de agrupamento. [32]

Avaliação e avaliação

A avaliação (ou "validação") dos resultados do clustering é tão difícil quanto o próprio clustering. [33] Abordagens populares envolvem avaliação " *interna* ", onde o agrupamento é resumido em uma única pontuação de qualidade, avaliação " *externa* ", onde o agrupamento é comparado a uma classificação existente de "verdade do terreno", avaliação " *manual* " por um especialista humano, e avaliação " *indireta* " avaliando a utilidade do agrupamento em sua aplicação pretendida. [34]

As medidas de avaliação interna têm o problema de representar funções que podem ser vistas como um objetivo de agrupamento. Por exemplo, pode-se agrupar os dados definidos pelo coeficiente Silhouette; exceto que não há algoritmo eficiente conhecido para isso. Ao usar tal medida interna para avaliação, pode-se antes comparar a similaridade dos problemas de otimização, [34] e não necessariamente o quão útil é o agrupamento.

A avaliação externa tem problemas semelhantes: se tivermos esses rótulos de "verdade fundamental", não precisaríamos nos agrupar; e em aplicações práticas geralmente não temos esses rótulos. Por outro lado, os rótulos refletem apenas um possível particionamento do conjunto de dados, o que não significa que não exista um clustering diferente, e talvez até melhor.

Nenhuma dessas abordagens pode, portanto, em última análise, julgar a qualidade real de um agrupamento, mas isso requer avaliação humana, [34] que é altamente subjetiva. No entanto, tais estatísticas podem ser bastante informativas na identificação de agrupamentos ruins, [35] mas não se deve descartar a avaliação humana subjetiva. [35]

Avaliação interna

Quando um resultado de clustering é avaliado com base nos dados que foram armazenados em cluster, isso é chamado de avaliação interna. Esses métodos geralmente atribuem a melhor pontuação ao algoritmo que produz clusters com alta similaridade dentro de um cluster e baixa similaridade entre os clusters. Uma desvantagem de usar critérios internos na avaliação de cluster é que pontuações altas em uma medida interna não resultam necessariamente em aplicativos eficazes de recuperação de informações. [36] Além disso, esta avaliação é tendenciosa para algoritmos que usam o mesmo modelo de cluster. Por exemplo, o agrupamento k-means otimiza naturalmente as distâncias dos objetos e um critério interno baseado na distância provavelmente superestima o agrupamento resultante.

Portanto, as medidas de avaliação interna são mais adequadas para obter alguns insights sobre situações em que um algoritmo tem desempenho melhor do que outro, mas isso não deve implicar que um algoritmo produza mais resultados válidos do que outro. [5] A validade medida por tal índice depende da afirmação de que este tipo de estrutura existe no conjunto de dados. Um algoritmo projetado para algum tipo de modelo não tem chance se o conjunto de dados contém um conjunto radicalmente diferente de modelos, ou se a avaliação mede um critério radicalmente diferente. [5] Por exemplo, o agrupamento k-means só pode encontrar clusters convexos, e muitos índices de avaliação assumem clusters convexos. Em um conjunto de dados com clusters não convexos, nem o uso de *k* -means, nem de um critério de avaliação que assume convexidade, é válido.

Existem mais de uma dúzia de medidas de avaliação interna, geralmente baseadas na intuição de que itens no mesmo cluster devem ser mais semelhantes do que itens em diferentes clusters. [37] : 115-121 Por exemplo, os seguintes métodos podem ser usados ​​para avaliar a qualidade dos algoritmos de agrupamento com base no critério interno:

* **Índice Davies-Bouldin**

O índice Davies-Bouldin pode ser calculado pela seguinte fórmula:

onde *n* é o número de clusters, é o centroide do cluster , é a distância média de todos os elementos no cluster para centróide , e é a distância entre os centróides e . Uma vez que algoritmos que produzem clusters com baixas distâncias intra-cluster (alta similaridade intra-cluster) e altas distâncias inter-cluster (baixa similaridade inter-cluster) terão um baixo índice de Davies – Bouldin, o algoritmo de agrupamento que produz uma coleção de clusters com o menor índice Davies – Bouldin é considerado o melhor algoritmo com base neste critério.

* **Índice de Dunn**

O índice de Dunn visa identificar clusters densos e bem separados. É definido como a razão entre a distância mínima entre clusters e a distância máxima entre clusters. Para cada partição do cluster, o índice de Dunn pode ser calculado pela seguinte fórmula: [38]

onde *d* ( *i* , *j* ) representa a distância entre os clusters *i* e *j* , e *d* '( *k* ) mede a distância intra-cluster do cluster *k* . A distância intercluster *d* ( *i* , *j* ) entre dois aglomerados pode ser qualquer número de medidas de distância, como a distância entre os centróides dos aglomerados. Da mesma forma, a distância intra-cluster *d* '( *k* ) pode ser medida de várias maneiras, como a distância máxima entre qualquer par de elementos no cluster *k* . Como o critério interno busca clusters com alta similaridade intra-cluster e baixa similaridade inter-cluster, algoritmos que produzem clusters com alto índice de Dunn são mais desejáveis.

* **Coeficiente de silhueta**

O coeficiente de silhueta contrasta a distância média para elementos no mesmo cluster com a distância média para elementos em outros clusters. Objetos com alto valor de silhueta são considerados bem agrupados, objetos com baixo valor podem ser outliers. Este índice funciona bem com agrupamento *k* -means, [ *carece de fontes* ] e também é usado para determinar o número ideal de clusters.

Avaliação externa

Na avaliação externa, os resultados do agrupamento são avaliados com base em dados que não foram usados ​​para agrupamento, como rótulos de classe conhecidos e benchmarks externos. Esses benchmarks consistem em um conjunto de itens pré-classificados, e esses conjuntos são frequentemente criados por humanos (especialistas). Assim, os conjuntos de referência podem ser considerados um padrão ouro para avaliação. [33] Esses tipos de métodos de avaliação medem o quão próximo o agrupamento está das classes de benchmark predeterminadas. No entanto, recentemente foi discutido se isso é adequado para dados reais, ou apenas em conjuntos de dados sintéticos com uma verdade fundamental factual, uma vez que as classes podem conter estrutura interna, os atributos presentes podem não permitir a separação de clusters ou as classes podem conter anomalias . [39] Além disso, do ponto de vista da descoberta do conhecimento , a reprodução do conhecimento conhecido pode não ser necessariamente o resultado pretendido. [39] No cenário especial de agrupamento restrito , onde meta informação (como rótulos de classe) já é usada no processo de agrupamento, a retenção de informações para fins de avaliação não é trivial. [40]

Uma série de medidas são adaptadas de variantes usadas para avaliar tarefas de classificação. Em vez de contar o número de vezes que uma classe foi atribuída corretamente a um único ponto de dados (conhecidos como verdadeiros positivos ), essas métricas de *contagem de pares* avaliam se cada par de pontos de dados que está realmente no mesmo cluster está previsto para estar no mesmo grupo. [33]

Tal como acontece com a avaliação interna, existem várias medidas de avaliação externa, [37] : 125–129 por exemplo:

* **Pureza** : Pureza é uma medida da extensão em que os clusters contêm uma única classe. [36] Seu cálculo pode ser pensado da seguinte maneira: Para cada cluster, conte o número de pontos de dados da classe mais comum no referido cluster. Agora pegue a soma de todos os clusters e divida pelo número total de pontos de dados. Formalmente, dado algum conjunto de clusters e algum conjunto de classes , ambos particionamento pontos de dados, a pureza pode ser definida como:

Esta medida não penaliza a existência de muitos clusters, e mais clusters tornará mais fácil a produção de alta pureza. Uma pontuação de pureza de 1 sempre é possível colocando cada ponto de dados em seu próprio cluster. Além disso, a pureza não funciona bem para dados desequilibrados, em que mesmo algoritmos de agrupamento de baixo desempenho fornecem um alto valor de pureza. Por exemplo, se um conjunto de dados de tamanho 1000 consiste em duas classes, uma contendo 999 pontos e a outra contendo 1 ponto, então cada partição possível terá uma pureza de pelo menos 99,9%.

* **Índice de Rand** [41]

O índice Rand calcula a semelhança dos clusters (retornados pelo algoritmo de clustering) com as classificações de benchmark. Ele pode ser calculado usando a seguinte fórmula:

Onde é o número de verdadeiros positivos, é o número de verdadeiros negativos , é o número de falsos positivos , e é o número de falsos negativos . As instâncias sendo contadas aqui são o número de atribuições de *pares* corretas . Isso é, é o número de pares de pontos que estão agrupados na partição prevista e na partição de verdade fundamental, é o número de pares de pontos que estão agrupados na partição prevista, mas não na partição verdadeira etc. Se o conjunto de dados for de tamanho N, então .

Um problema com o índice de Rand é que falsos positivos e falsos negativos são igualmente ponderados. Esta pode ser uma característica indesejável para alguns aplicativos de cluster. A medida F aborda essa preocupação, [ *carece de fontes?* ], Assim como o índice de Rand ajustado com correção de chance .

* **Medida F**

A medida F pode ser usada para equilibrar a contribuição de falsos negativos , ponderando a recuperação por meio de um parâmetro . Deixe que a **precisão** e a **recuperação** (ambas as medidas de avaliação externa em si) sejam definidas da seguinte forma:

Onde é a taxa de precisão e é a taxa de recall . Podemos calcular a medida F usando a seguinte fórmula: [36]

Quando , . Em outras palavras, o recall não tem impacto sobre a medida F quando , e aumentando aloca uma quantidade crescente de peso para lembrar no compasso F final.

Também não é levado em consideração e pode variar de 0 para cima sem limite.

* **Índice de Jaccard**

O índice de Jaccard é usado para quantificar a similaridade entre dois conjuntos de dados. O índice Jaccard assume um valor entre 0 e 1. Um índice 1 significa que os dois conjuntos de dados são idênticos e um índice 0 indica que os conjuntos de dados não têm elementos comuns. O índice Jaccard é definido pela seguinte fórmula:

Este é simplesmente o número de elementos exclusivos comuns a ambos os conjuntos dividido pelo número total de elementos exclusivos em ambos os conjuntos.

Também não é levado em consideração e pode variar de 0 para cima sem limite.

* **Índice de dados**

A medida simétrica Dice dobra o peso em enquanto ainda ignora :

* **Índice de Fowlkes-Mallows** [42]

O índice de Fowlkes – Mallows calcula a similaridade entre os clusters retornados pelo algoritmo de clusterização e as classificações de benchmark. Quanto mais alto o valor do índice de Fowlkes-Mallows, mais semelhantes são os clusters e as classificações de benchmark. Ele pode ser calculado usando a seguinte fórmula:

Onde é o número de verdadeiros positivos , é o número de falsos positivos , e é o número de falsos negativos . O índice é a média geométrica da precisão e recuperação e , e, portanto, também é conhecido como a medida G, enquanto a medida F é sua média harmônica. [43] [44] Além disso, precisão e recall também são conhecidos como índices de Wallace e . [45] Versões normalizadas de chance de recall, precisão e medida-G correspondem a Informedness , Markedness e Matthews Correlation e estão fortemente relacionadas ao Kappa . [46]

* A **informação mútua** é uma medida teórica da informação de quanta informação é compartilhada entre um agrupamento e uma classificação de verdade do solo que pode detectar uma similaridade não linear entre dois agrupamentos. A informação mútua normalizada é uma família de variantes corrigidas para o acaso, que tem um viés reduzido para vários números de cluster. [33]
* **Matriz de confusão**

Uma matriz de confusão pode ser usada para visualizar rapidamente os resultados de um algoritmo de classificação (ou agrupamento). Mostra como um cluster é diferente do cluster padrão ouro.

Tendência de cluster

Medir a tendência do cluster é medir em que grau os clusters existem nos dados a serem agrupados e pode ser realizado como um teste inicial, antes de tentar o agrupamento. Uma maneira de fazer isso é comparar os dados com os dados aleatórios. Em média, os dados aleatórios não devem ter clusters.

* **Estatística de Hopkins**

Existem várias formulações da estatística de Hopkins . [47] Um típico é o seguinte. [48] Let seja o conjunto de pontos de dados em espaço dimensional. Considere uma amostra aleatória (sem substituição) de pontos de dados com membros . Também gere um conjunto de pontos de dados uniformemente distribuídos aleatoriamente. Agora defina duas medidas de distância, estar a distância de de seu vizinho mais próximo em X e estar a distância de de seu vizinho mais próximo em X. Em seguida, definimos a estatística de Hopkins como:

Com essa definição, dados aleatórios uniformes tendem a ter valores próximos a 0,5, e dados agrupados tendem a ter valores próximos a 1.

No entanto, dados contendo apenas um Gaussiano também pontuam perto de 1, pois essa estatística mede o desvio de uma distribuição *uniforme* , não a multimodalidade , tornando essa estatística amplamente inútil na aplicação (já que os dados reais nunca são remotamente uniformes).

Formulários

Biologia, biologia computacional e bioinformática

**Ecologia vegetal e animal**

A análise de agrupamento é usada para descrever e fazer comparações espaciais e temporais de comunidades (assembléias) de organismos em ambientes heterogêneos. Também é usado na sistemática de plantas para gerar filogenias artificiais ou agrupamentos de organismos (indivíduos) na espécie, gênero ou nível superior que compartilham uma série de atributos.

**Transcriptômica**

O agrupamento é usado para construir grupos de genes com padrões de expressão relacionados (também conhecidos como genes coexpressos) como no algoritmo de agrupamento HCS . [49] [50] Freqüentemente, esses grupos contêm proteínas funcionalmente relacionadas, como enzimas para uma via específica ou genes que são co-regulados. Experimentos de alto rendimento usando tags de sequência expressa (ESTs) ou microarranjos de DNA podem ser uma ferramenta poderosa para anotação de genoma - um aspecto geral da genômica .

**Análise de sequência**

O agrupamento de sequências é usado para agrupar sequências homólogas em famílias de genes . [51] Este é um conceito muito importante em bioinformática e biologia evolutiva em geral. Veja a evolução por duplicação de genes .

**Plataformas de genotipagem de alto rendimento**

Os algoritmos de agrupamento são usados ​​para atribuir genótipos automaticamente. [52]

**Agrupamento genético humano**

A similaridade de dados genéticos é usada em agrupamento para inferir estruturas populacionais.

Medicamento

**Imagens médicas**

Em exames PET , a análise de cluster pode ser usada para diferenciar entre diferentes tipos de tecido em uma imagem tridimensional para muitos propósitos diferentes. [53]

**Análise da atividade antimicrobiana**

A análise de agrupamento pode ser usada para analisar os padrões de resistência aos antibióticos, para classificar os compostos antimicrobianos de acordo com seu mecanismo de ação, para classificar os antibióticos de acordo com sua atividade antibacteriana.

**Segmentação IMRT**

O agrupamento pode ser usado para dividir um mapa de fluência em regiões distintas para conversão em campos de entrega em terapia de radiação baseada em MLC.

Negócios e marketing

**Pesquisa de mercado**

A análise de cluster é amplamente usada em pesquisas de mercado ao trabalhar com dados multivariados de pesquisas e painéis de teste. Os pesquisadores de mercado usam a análise de cluster para dividir a população geral de consumidores em segmentos de mercado e para entender melhor as relações entre diferentes grupos de consumidores / clientes potenciais e para uso em segmentação de mercado , posicionamento de produto , desenvolvimento de novos produtos e seleção de mercados de teste.

**Agrupamento de itens de compras**

O clustering pode ser usado para agrupar todos os itens de compras disponíveis na web em um conjunto de produtos exclusivos. Por exemplo, todos os itens no eBay podem ser agrupados em produtos exclusivos (o eBay não tem o conceito de SKU ).

Rede mundial de computadores

**Análise de rede social**

No estudo de redes sociais , o agrupamento pode ser usado para reconhecer comunidades dentro de grandes grupos de pessoas.

**Agrupamento de resultados de pesquisa**

No processo de agrupamento inteligente dos arquivos e sites, o agrupamento pode ser usado para criar um conjunto mais relevante de resultados de pesquisa em comparação com motores de pesquisa normais como o Google [ *carece de fontes?* ] . Atualmente, há uma série de ferramentas de cluster baseadas na web, como o Clusty . Ele também pode ser usado para retornar um conjunto mais abrangente de resultados nos casos em que um termo de pesquisa pode se referir a coisas muito diferentes. Cada uso distinto do termo corresponde a um cluster exclusivo de resultados, permitindo que um algoritmo de classificação retorne resultados abrangentes selecionando o resultado principal de cada cluster. [54]

**Otimização de mapa deslizante**

O mapa de fotos do Flickr e outros sites de mapas usam agrupamento para reduzir o número de marcadores em um mapa. Isso o torna mais rápido e reduz a quantidade de confusão visual.

Ciência da Computação

**Evolução de software**

O armazenamento em cluster é útil na evolução do software, pois ajuda a reduzir as propriedades legadas no código, reformando a funcionalidade que se tornou dispersa. É uma forma de reestruturação e, portanto, uma forma de manutenção preventiva direta.

**Segmentação de imagem**

O agrupamento pode ser usado para dividir uma imagem digital em regiões distintas para detecção de bordas ou reconhecimento de objetos . [55]

**Algoritmos evolutivos**

O agrupamento pode ser usado para identificar diferentes nichos dentro da população de um algoritmo evolutivo, de modo que a oportunidade reprodutiva possa ser distribuída de maneira mais uniforme entre as espécies ou subespécies em evolução.

**Sistemas de recomendação**

Os sistemas de recomendação são projetados para recomendar novos itens com base nas preferências do usuário. Eles às vezes usam algoritmos de cluster para prever as preferências de um usuário com base nas preferências de outros usuários no cluster do usuário.

**Métodos de Monte Carlo da cadeia de Markov**

O agrupamento é frequentemente utilizado para localizar e caracterizar extremos na distribuição de destino.

**Detecção de anomalia**

Anomalias / outliers são tipicamente - seja explicitamente ou implicitamente - definidos em relação à estrutura de agrupamento de dados.

**Processamento de linguagem natural**

O agrupamento pode ser usado para resolver a ambigüidade lexical . [54]

Ciências Sociais

**Análise de crime**

A análise de agrupamento pode ser usada para identificar áreas onde há maior incidência de tipos específicos de crime. Ao identificar essas áreas distintas ou "pontos críticos" onde um crime semelhante aconteceu durante um período de tempo, é possível gerenciar os recursos de aplicação da lei de forma mais eficaz.

**Mineração de dados educacionais**

A análise de cluster é, por exemplo, usada para identificar grupos de escolas ou alunos com propriedades semelhantes.

**Tipologias**

A partir de dados de pesquisas, projetos como os realizados pelo Pew Research Center usam análise de cluster para discernir tipologias de opiniões, hábitos e dados demográficos que podem ser úteis em política e marketing.

Outras

**Robótica de campo**

Algoritmos de clustering são usados ​​para consciência situacional robótica para rastrear objetos e detectar outliers nos dados do sensor. [56]

**Química matemática**

Para encontrar semelhanças estruturais, etc., por exemplo, 3.000 compostos químicos foram agrupados no espaço de 90 índices topológicos . [57]

**Climatologia**

Para encontrar regimes meteorológicos ou padrões atmosféricos preferenciais de pressão ao nível do mar. [58]

**Finança**

A análise de cluster foi usada para agrupar ações em setores. [59]

**Geologia do petróleo**

A análise de cluster é usada para reconstruir dados de núcleo de fundo de poço ausentes ou curvas de registro ausentes, a fim de avaliar as propriedades do reservatório.

**Geoquímica**

O agrupamento de propriedades químicas em diferentes locais de amostra.

Veja também

|  |
| --- |
|  |

Tipos especializados de análise de cluster

* Algoritmos de clustering automático
* Clustering balanceado
* Agrupando dados de alta dimensão
* Agrupamento conceitual
* Agrupamento de consenso
* Clustering restrito
* Detecção de comunidade
* Agrupamento de fluxo de dados
* Clustering HCS
* Agrupamento de sequência
* Agrupamento espectral

Técnicas usadas na análise de cluster

* Rede neural artificial (ANN)
* Pesquisa de vizinho mais próximo
* Análise de componentes de vizinhança
* Análise de classe latente
* Propagação de afinidade

Projeção e pré-processamento de dados

* Redução de dimensão
* Análise do componente principal
* Escala multidimensional

Outro

* Modelagem ponderada por cluster
* Maldição da dimensionalidade
* Determinar o número de clusters em um conjunto de dados
* Coordenadas paralelas
* Análise de dados estruturados

Referências

1. **^** *Driver e Kroeber (1932). "Expressão Quantitativa das Relações Culturais" . Publicações da Universidade da Califórnia em Arqueologia e Etnologia Americanas . Quantitative Expression of Cultural Relationships: 211–256 - via http://dpg.lib.berkeley.edu .*
2. **^** *Zubin, Joseph (1938). "Uma técnica para medir pensamentos semelhantes". The Journal of Abnormal and Social Psychology .* ***33*** *(4): 508–516. doi : 10.1037 / h0055441 . ISSN 0096-851X .*
3. **^** *Tryon, Robert C. (1939). Análise de Cluster: Perfil de Correlação e Análise Ortométrica (fator) para o Isolamento de Unidades na Mente e Personalidade . Irmãos Edwards.*
4. **^** *Cattell, RB (1943). "A descrição da personalidade: traços básicos resolvidos em grupos". Journal of Abnormal and Social Psychology .* ***38*** *(4): 476–506. doi : 10.1037 / h0054116 .*
5. ^ a b c d e f *Estivill-Castro, Vladimir (20 de junho de 2002). "Por que tantos algoritmos de agrupamento - Um documento de posição". Boletim Informativo ACM SIGKDD Explorations .* ***4*** *(1): 65–75. doi : 10.1145 / 568574.568575 . S2CID 7329935 .*
6. **^** James A. Davis (maio de 1967) "Clustering and structure balance in graphs", Human Relations 20: 181-7
7. **^** *Everitt, Brian (2011). Análise de cluster . Chichester, West Sussex, Reino Unido: Wiley. ISBN 9780470749913.*
8. **^** *Sibson, R. (1973). "SLINK: um algoritmo de eficiência otimizada para o método de cluster de link único" (PDF) . The Computer Journal . Sociedade Britânica de Computação.* ***16*** *(1): 30–34. doi : 10.1093 / comjnl / 16.1.30 .*
9. **^** *Defays, D. (1977). "Um algoritmo eficiente para um método de link completo". The Computer Journal . Sociedade Britânica de Computação.* ***20*** *(4): 364–366. doi : 10.1093 / comjnl / 20.4.364 .*
10. **^** *Lloyd, S. (1982). "Quantização de mínimos quadrados em PCM". IEEE Transactions on Information Theory .* ***28*** *(2): 129–137. doi : 10.1109 / TIT.1982.1056489 .*
11. **^** *Kriegel, Hans-Peter ; Kröger, Peer; Sander, Jörg; Zimek, Arthur (2011). "Clustering baseado em densidade" . WIREs Data Mining and Knowledge Discovery .* ***1*** *(3): 231–240. doi : 10.1002 / widm.30 . S2CID 36920706 .*
12. **^** Pesquisa acadêmica da Microsoft: artigos de mineração de dados mais citados. Arquivado em21/04/2010na Wayback Machine : DBSCAN está no rank 24, quando acessado em: 18/04/2010
13. **^** *Ester, Martin; Kriegel, Hans-Peter ; Sander, Jörg; Xu, Xiaowei (1996). "Um algoritmo baseado em densidade para descobrir clusters em grandes bancos de dados espaciais com ruído". Em Simoudis, Evangelos; Han, Jiawei; Fayyad, Usama M. (eds.). Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96) . AAAI Press . pp. 226–231. ISBN 1-57735-004-9.*
14. **^** *Ankerst, Mihael; Breunig, Markus M .; Kriegel, Hans-Peter ; Sander, Jörg (1999). "ÓPTICA: Pontos de pedido para identificar a estrutura de agrupamento". ACM SIGMOD conferência internacional sobre gestão de dados . ACM Press . pp. 49–60. CiteSeerX 10.1.1.129.6542 .*
15. ^ a b *Achtert, E .; Bohm, C .; Kröger, P. (2006). "DeLi-Clu: Aumentando a robustez, integridade, usabilidade e eficiência do agrupamento hierárquico por uma classificação de par mais próximo". Avanços na descoberta de conhecimento e mineração de dados . Notas de aula em Ciência da Computação.* ***3918*** *. pp. 119–128. CiteSeerX 10.1.1.64.1161 . doi : 10.1007 / 11731139\_16 . ISBN 978-3-540-33206-0.*
16. **^** *Aggarwal, Charu C., editor. Reddy, Chandan K., editor. Clustering de dados: algoritmos e aplicações . ISBN 978-1-315-37351-5. OCLC 1110589522 .*CS1 maint: vários nomes: lista de autores ( link )
17. **^** *Sculley, D. (2010). Clustering k-means na escala da Web . Proc. 19º WWW.*
18. **^** *Huang, Z. (1998). "Extensões ao algoritmo k- significa para agrupar grandes conjuntos de dados com valores categóricos". Mineração de dados e descoberta de conhecimento .* ***2*** *(3): 283–304. doi : 10.1023 / A: 1009769707641 . S2CID 11323096 .*
19. **^** R. Ng e J. Han. "Método de agrupamento eficiente e eficaz para mineração de dados espaciais". In: Anais da 20ª Conferência VLDB, páginas 144–155, Santiago, Chile, 1994.
20. **^** Tian Zhang, Raghu Ramakrishnan, Miron Livny. " Um método de agrupamento de dados eficiente para bancos de dados muito grandes ." In: Proc. Int'l Conf. on Management of Data, ACM SIGMOD, pp. 103-114.
21. **^** *Kriegel, Hans-Peter ; Kröger, Peer; Zimek, Arthur (julho de 2012). "Clustering de subespaço". Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery .* ***2*** *(4): 351–364. doi : 10.1002 / widm.1057 . S2CID 7241355 .*
22. **^** *Agrawal, R .; Gehrke, J .; Gunopulos, D .; Raghavan, P. (2005). "Clustering Subespacial Automático de Dados Dimensionais Elevados". Mineração de dados e descoberta de conhecimento .* ***11*** *: 5-33. CiteSeerX 10.1.1.131.5152 . doi : 10.1007 / s10618-005-1396-1 . S2CID 9289572 .*
23. **^** Karin Kailing, Hans-Peter Kriegel e Peer Kröger. *Clustering de subespaço conectado por densidade para dados de alta dimensão* . In: *Proc. SIAM Int. Conf. on Data Mining (SDM'04)* , pp. 246-257, 2004.
24. **^** *Achtert, E .; Bohm, C .; Kriegel, H.-P. ; Kröger, P .; Müller-Gorman, I .; Zimek, A. (2006). "Encontrando Hierarquias de Clusters de Subspace". Descoberta de conhecimento em bancos de dados: PKDD 2006 . Notas de aula em Ciência da Computação.* ***4213*** *. pp. 446–453. CiteSeerX 10.1.1.705.2956 . doi : 10.1007 / 11871637\_42 . ISBN 978-3-540-45374-1.*
25. **^** *Achtert, E .; Bohm, C .; Kriegel, HP ; Kröger, P .; Müller-Gorman, I .; Zimek, A. (2007). "Detecção e visualização de hierarquias de cluster de subespaço". Avanços em Bancos de Dados: Conceitos, Sistemas e Aplicações . Notas de aula em Ciência da Computação.* ***4443*** *. pp. 152–163. CiteSeerX 10.1.1.70.7843 . doi : 10.1007 / 978-3-540-71703-4\_15 . ISBN 978-3-540-71702-7.*
26. **^** *Achtert, E .; Bohm, C .; Kröger, P .; Zimek, A. (2006). "Hierarquias de mineração de clusters de correlação". Proc. 18ª Conferência Internacional sobre Gerenciamento de Banco de Dados Científico e Estatístico (SSDBM) : 119–128. CiteSeerX 10.1.1.707.7872 . doi : 10.1109 / SSDBM.2006.35 . ISBN 978-0-7695-2590-7. S2CID 2679909 .*
27. **^** *Bohm, C .; Kailing, K .; Kröger, P .; Zimek, A. (2004). "Computando Clusters de Objetos Conectados por Correlação". Anais da conferência internacional ACM SIGMOD 2004 sobre gerenciamento de dados - SIGMOD '04 . p. 455. CiteSeerX 10.1.1.5.1279 . doi : 10.1145 / 1007568.1007620 . ISBN 978-1581138597. S2CID 6411037 .*
28. **^** *Achtert, E .; Bohm, C .; Kriegel, HP ; Kröger, P .; Zimek, A. (2007). "Explorando Relacionamentos Complexos de Clusters de Correlação". 19ª Conferência Internacional sobre Gerenciamento de Banco de Dados Científico e Estatístico (SSDBM 2007) . p. 7. CiteSeerX 10.1.1.71.5021 . doi : 10.1109 / SSDBM.2007.21 . ISBN 978-0-7695-2868-7. S2CID 1554722 .*
29. **^** *Meilă, Marina (2003). "Comparando Clusterings pela Variação da Informação". Teoria de Aprendizagem e Máquinas Kernel . Notas de aula em Ciência da Computação.* ***2777*** *. pp. 173–187. doi : 10.1007 / 978-3-540-45167-9\_14 . ISBN 978-3-540-40720-1.*
30. **^** *Kraskov, Alexander; Stögbauer, Harald; Andrzejak, Ralph G .; Grassberger, Peter (1 de dezembro de 2003). "Clustering hierárquico baseado em informações mútuas". arXiv : q-bio / 0311039 . Bibcode : 2003q.bio .... 11039K .* Citar diário requer |journal=( ajuda )
31. **^** *Auffarth, B. (18–23 de julho de 2010). "Agrupamento por um algoritmo genético com operador de mutação enviesado" . Wcci Cec . IEEE.*
32. **^** *Frey, BJ; Dueck, D. (2007). "Clustering pela passagem de mensagens entre pontos de dados". Ciência .* ***315*** *(5814): 972–976. Bibcode : 2007Sci ... 315..972F . CiteSeerX 10.1.1.121.3145 . doi : 10.1126 / science.1136800 . PMID 17218491 . S2CID 6502291 .*
33. ^ a b c d *Pfitzner, Darius; Leibbrandt, Richard; Powers, David (2009). "Caracterização e avaliação de medidas de similaridade para pares de agrupamentos". Conhecimento e Sistemas de Informação . Springer.* ***19*** *(3): 361–394. doi : 10.1007 / s10115-008-0150-6 . S2CID 6935380 .*
34. ^ a b c *Feldman, Ronen; Sanger, James (01-01-2007). The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data . Cambridge Univ. Aperte. ISBN 978-0521836579. OCLC 915286380 .*
35. ^ a b *Weiss, Sholom M .; Indurkhya, Nitin; Zhang, Tong; Damerau, Fred J. (2005). Mineração de texto: métodos preditivos para analisar informações não estruturadas . Springer. ISBN 978-0387954332. OCLC 803401334 .*
36. ^ a b c *Manning, Christopher D .; Raghavan, Prabhakar; Schütze, Hinrich (07/07/2008). Introdução à recuperação de informações . Cambridge University Press. ISBN 978-0-521-86571-5.*
37. ^ a b *Knowledge Discovery in Databases - Part III - Clustering (PDF) , Heidelberg University , 2017*
38. **^** *Dunn, J. (1974). "Clusters bem separados e partições fuzzy ideais". Journal of Cybernetics .* ***4*** *: 95–104. doi : 10.1080 / 01969727408546059 .*
39. ^ a b *Färber, Ines; Günnemann, Stephan; Kriegel, Hans-Peter ; Kröger, Peer; Müller, Emmanuel; Schubert, Erich; Seidl, Thomas; Zimek, Arthur (2010). "On Using Class-Labels in Evaluation of Clusterings" (PDF) . Em Fern, Xiaoli Z .; Davidson, Ian; Dy, Jennifer (eds.). MultiClust: Descobrindo, resumindo e usando vários agrupamentos . ACM SIGKDD .*
40. **^** *Pourrajabi, M .; Moulavi, D .; Campello, RJGB; Zimek, A .; Sander, J .; Goebel, R. (2014). "Seleção de modelo para agrupamento semi-supervisionado". Proceedings of the 17th International Conference on Extending Database Technology (EDBT) . pp. 331–342. doi : 10.5441 / 002 / edbt.2014.31 .*
41. **^** *Rand, WM (1971). “Critérios objetivos para a avaliação dos métodos de agrupamento”. Journal of the American Statistical Association . American Statistical Association.* ***66*** *(336): 846–850. arXiv : 1704.01036 . doi : 10.2307 / 2284239 . JSTOR 2284239 .*
42. **^** *Fowlkes, EB; Mallows, CL (1983). "Um método para comparar dois agrupamentos hierárquicos". Journal of the American Statistical Association .* ***78*** *(383): 553–569. doi : 10.1080 / 01621459.1983.10478008 . JSTOR 2288117 .*
43. **^** *Powers, David (2003). Recall e precisão versus o bookmaker . Conferência Internacional sobre Ciência Cognitiva. pp. 529–534.*
44. **^** *Arabie, P. (1985). "Comparando partições". Journal of Classification .* ***2*** *(1): 1985. doi : 10.1007 / BF01908075 . S2CID 189915041 .*
45. **^** *Wallace, DL (1983). "Comente". Journal of the American Statistical Association .* ***78*** *(383): 569–579. doi : 10.1080 / 01621459.1983.10478009 .*
46. **^** *Powers, David (2012). O problema com Kappa . Capítulo Europeu da Association for Computational Linguistics. pp. 345–355.*
47. **^** *Hopkins, Brian; Skellam, John Gordon (1954). “Um novo método para determinar o tipo de distribuição dos indivíduos das plantas”. Annals of Botany . Annals Botany Co.* ***18*** *(2): 213–227. doi : 10.1093 / oxfordjournals.aob.a083391 .*
48. **^** *Banerjee, A. (2004). "Validando clusters usando a estatística de Hopkins". IEEE International Conference on Fuzzy Systems .* ***1*** *: 149–153. doi : 10.1109 / FUZZY.2004.1375706 . ISBN 978-0-7803-8353-1. S2CID 36701919 .*
49. **^** *Johnson, Stephen C. (01/09/1967). "Esquemas de agrupamento hierárquico". Psychometrika .* ***32*** *(3): 241–254. doi : 10.1007 / BF02289588 . ISSN 1860-0980 . PMID 5234703 . S2CID 930698 .*
50. **^** *Hartuv, Erez; Shamir, Ron (2000-12-31). “Um algoritmo de agrupamento baseado na conectividade do gráfico”. Cartas de processamento de informações .* ***76*** *(4): 175–181. doi : 10.1016 / S0020-0190 (00) 00142-3 . ISSN 0020-0190 .*
51. **^** *Remm, Maido; Tempestade, Christian EV; Sonnhammer, Erik LL (2001-12-14). "Agrupamento automático de ortólogos e em-parálogos de comparações de espécies em pares11Editado por F. Cohen". Journal of Molecular Biology .* ***314*** *(5): 1041–1052. doi : 10.1006 / jmbi.2000.5197 . ISSN 0022-2836 . PMID 11743721 .*
52. **^** *Botstein, David; Cox, David R .; Risch, Neil; Olshen, Richard; Curb, David; Dzau, Victor J .; Chen, Yii-Der I .; Hebert, Joan; Pesich, Robert (01/07/2001). "Genotipagem de alto rendimento com polimorfismos de nucleotídeo único" . Genome Research .* ***11*** *(7): 1262–1268. doi : 10.1101 / gr.157801 (inativo em 31 de maio de 2021). ISSN 1088-9051 . PMC 311112 . PMID 11435409 .*Manutenção CS1: DOI inativo em maio de 2021 ( link )
53. **^** *Filipovych, Roman; Resnick, Susan M .; Davatzikos, Christos (2011). "Análise de Cluster Semi-supervisionada de Dados de Imagem" . NeuroImage .* ***54*** *(3): 2185–2197. doi : 10.1016 / j.neuroimage.2010.09.074 . PMC 3008313 . PMID 20933091 .*
54. ^ a b *Di Marco, Antonio; Navigli, Roberto (2013). "Agrupando e diversificando resultados de pesquisa na Web com indução de sentido de palavras baseada em gráfico". Lingüística Computacional .* ***39*** *(3): 709–754. doi : 10.1162 / COLI\_a\_00148 . S2CID 1775181 .*
55. **^** Bewley, A., & Upcroft, B. (2013). Vantagens de explorar a estrutura de projeção para segmentar nuvens de pontos 3D densas. Na Conferência Australiana sobre Robótica e Automação [1]
56. **^** *Bewley, A .; et al. "Estimativa de volume em tempo real de uma carga útil de dragline". Conferência Internacional IEEE sobre Robótica e Automação .* ***2011*** *: 1571–1576.*
57. **^** *Basak, SC; Magnuson, VR; Niemi, CJ; Regal, RR (1988). "Determinando a similaridade estrutural de produtos químicos usando índices teóricos de grafos". Discr. Appl. Matemática .* ***19*** *(1-3): 17-44. doi : 10.1016 / 0166-218x (88) 90004-2 .*
58. **^** *Huth, R .; et al. (2008). "Classificações de padrões de circulação atmosférica: avanços recentes e aplicações". Ann. NY Acad. Sci .* ***1146*** *: 105–152. Bibcode : 2008NYASA1146..105H . doi : 10.1196 / annals.1446.019 . PMID 19076414 . S2CID 22655306 .*
59. **^** *Arnott, Robert D. (01/11/1980). "Análise de Cluster e Comovement do Preço das Ações". Financial Analysts Journal .* ***36*** *(6): 56–62. doi : 10.2469 / faj.v36.n6.56 . ISSN 0015-198X .*