Técnicas de Agrupamento (Clustering)

Sarajane M. Peres e Clodoaldo A. M. Lima

17 de setembro de 2015

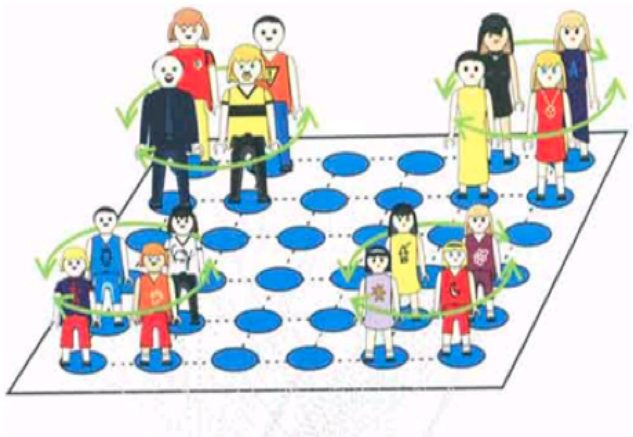
Material baseado em:
HAN, J. & KAMBER, M. Data Mining: Concepts and Techniques. 2nd. 2006
ROCHA, T., PERES, S. M., BÍSCARO, H. H., MADEO, R. C. B., BOSCARIOLI, C. Tutorial sobre Fuzzu-c-Means e Fuzzy Learning Vector Quantization: Abordagens Híbridas para Tarefas de Agrupamentos e Classificação. Revista de Informática Teórica e Aplicada, vol.9, n.1, 2012.
COSTA, J. A. F. Classificação Automática e Análise de Dados por Redes Neurais Auto-Organizáveis. Tese de Doutorado. Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas. 1999.

Clustering

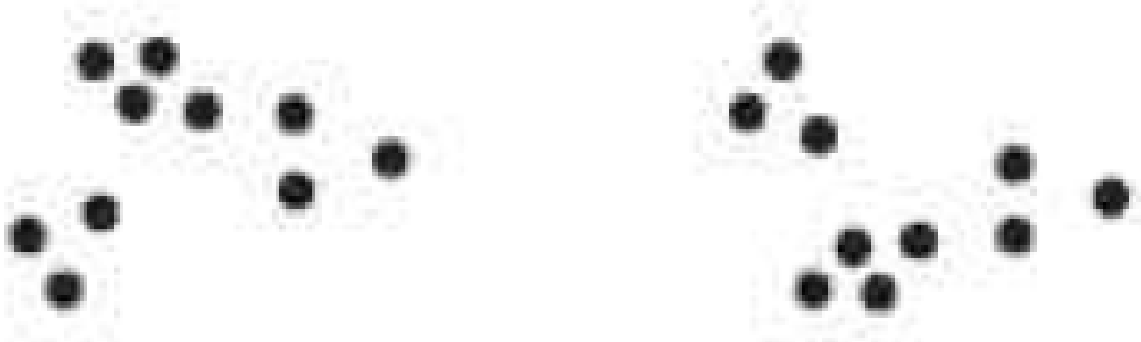
Clustering - Agrupamento

O termo *grupo* deve ser usado quando não existe qualquer informação sobre como é a organização dos dados. Nesse caso, o trabalho de análise de dados é denominado *agrupamento* (*clustering*), e tem por objetivo estudar as relações de similaridades entre os dados, determinando quais dados formam quais grupos.

Os grupos são formados de maneira a maximizar a similaridade entre os elementos de um grupo (similaridade intra-grupo) e minimizar a similaridade entre elementos de grupos diferentes (similaridade inter-grupos).



Clustering



(a) Original

Clustering

(a) Original

(b) Dois grupos

(c) Quatro grupos

(d) Seis grupos

Clustering

Para o contexto de nosso estudo, um conjunto de dados X é definido como:

$$X = \begin{matrix} \vec{x_1} \\ \vec{x_2} \\ \cdots \\ \vec{x_n} \end{matrix} = \begin{matrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{matrix}$$

onde $\vec{x_j}$ é um vetor de p coordenadas e n é o número de elementos do conjunto de dados. Cada vetor representa um dado desse conjunto e cada coordenada desse vetor representa um atributo descritivo do dado. O conjunto de dados X reside no espaço $\Re^p$, e este espaço é referenciado pelos algoritmos de análise de dados como "espaço dos dados", "espaço de entrada" ou "espaço vetorial".

Clustering

Formalmente, dado um conjunto de dados de entrada $\vec{x} \in \Re^p$, é encontrada uma função

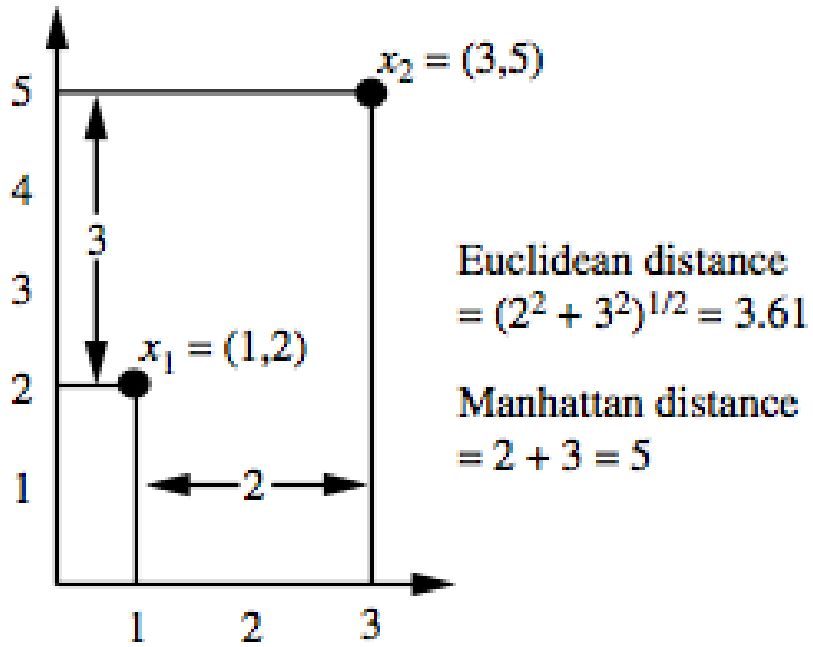
$$\mathcal{G} : \Re^p \times W \rightarrow C$$

onde W é um vetor de parâmetros ajustáveis, por meio de um algoritmo de aprendizado não supervisionado, que determina c-grupos em X, $C = C_1, ..., C_c (c \leq n)$ tal que:

- $C_i \neq \emptyset, i = 1, ..., c$;
- $\bigcup_{i=1}^{c} C_i = X$;
- $C_i \cap C_j = \emptyset, i, j = 1, ..., c$ and $i \neq j$, assumindo a abordagem de agrupamento clássica.

Clustering

Os algoritmos que executam tarefas de análise de dados, muitas vezes, usam alguma medida de similaridade entre vetores (dados) em seu processo de execução. Essas medidas servem para guiar o processo de construção da superfície de decisão que determinará qual é a região de abrangência de cada grupo de dados.



Clustering

Distância Euclidiana

$$d_{Euclidiana}(\vec{v}_i, \vec{v}_j) = \left(\sum_{l=1}^{p} (v_{il} - v_{jl})^2 \right)^{\frac{1}{2}}$$

onde p é a dimensão do espaço dos vetores e $\vec{v}_i$ e $\vec{v}_j$ são os vetores sobre os quais se deseja calcular a similaridade.

Distância Manhattan

$$d_{Manhattan}(\vec{v}_i, \vec{v}_j) = \sum_{l=1}^{p} |v_{il} - v_{jl}|$$

onde p é a dimensão do espaço dos vetores e $\vec{v}_i$ e $\vec{v}_j$ são os vetores sobre os quais se deseja calcular a similaridade.

Clustering

Categorização de Métodos de Clustering

- Método por particionamento: dado um conjunto de dados com n instâncias, um método por particionamento constrói k partições dos dados, onde cada partição representa um grupo e $k \leq n$. O método cria uma partição inicial e, então, usa uma técnica de realocação iterativa que tenta melhorar o particionamento. **Exemplo: c-Means (ou k-Means), CLARANS**.
- Métodos hierárquicos: cria uma decomposição hierárquica de um conjunto de dados. Os métodos hierárquicos podem ser *aglomerativos* ou *divisivos*, dependendo de como a decomposição hierárquica é formada - juntando decomposições ou dividindo composições. A cada passo, divisões ou junções são feitas. Podem representar seus resultados em dendogramas. **Exemplo: AGNES/DIANA, BIRCH, ROCK, Chameleon**.
- Métodos baseados em densidade: No caso dos métodos baseados em densidade, os grupos formados crescem de acordo com a densidade de dados em um "potencial" grupo. Para cada dado dentro de um dado grupo, a vizinhança em um dado raio tem que conter pelo menos um número mínimo de pontos. **Exemplo: DBSCAN, OPTICS, DENCLUE**.
- Métodos baseados em modelos: criam uma hipótese sobre um modelo para cada um dos grupos e encontram o melhor ajuste dos dados ao modelo. **Exemplo: Self-Organizing Map (SOM), Expectation-Maximization (EM)** .
- Métodos baseados em *grid*: esses métodos quantizam o espaço de dados em um número finito de células que forma uma estrutura em *grid*. **Exemplo: STING, WaveCluster**

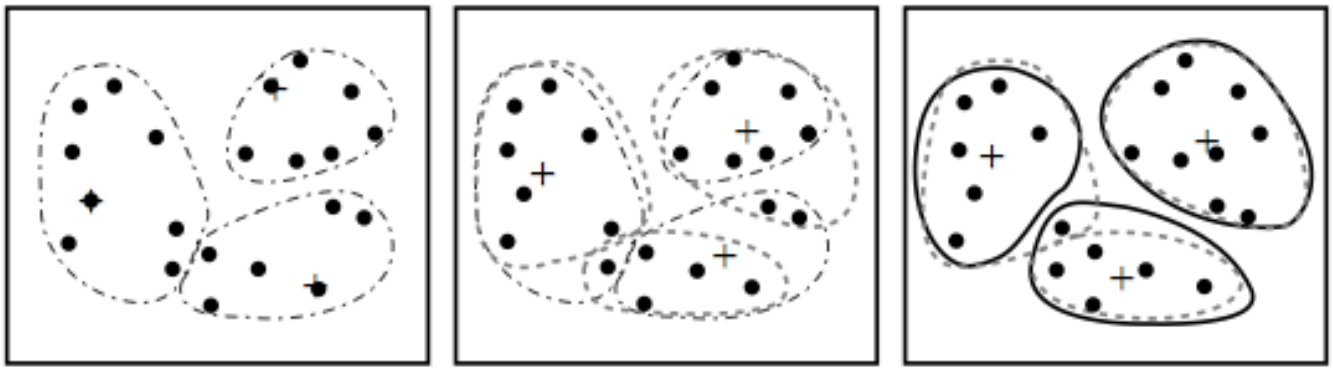
Clustering

Por particionamento ....

Clustering - Métodos por Particionamento

c-Means

Nesse algoritmo, c agrupamentos são representados como um conjunto $\mathcal{C} = \{\vec{C}_1, ..., \vec{C}_c\}$ de vetores chamados "protótipos". Cada vetor protótipo sempre está associado à representação de um grupo do conjunto de dados e, para isso, deve residir no mesmo espaço $\Re^p$ que os dados do conjunto. O conjunto $\mathcal{C}$ é representado por uma matriz de dimensão $c \times p$.



Clustering - Métodos por Particionamento

c-Means

Para alcançar seu objetivo, o algoritmo realiza várias iterações na busca de uma configuração ótima de parâmetros para minimizar $J_{CM}(U_h, C)$, que é dado por:

$$J_{CM}(U_h, \mathcal{C}) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij} d(\vec{C}_i, \vec{x}_j)^2 \quad (1)$$

onde $d(\vec{C}_i, \vec{x}_j)$, é a distância entre o vetor de dados $\vec{x}_j$ e o protótipo do grupo $\vec{C}_i$, c é o número de grupos a ser determinado pelo algoritmo, n é o número de dados no conjunto de dados e U_h é uma matriz binária chamada "matriz de partição", de dimensões $c \times n$, definida como:

$$U_h = \begin{bmatrix} u_{1,1} & u_{1,2} & \cdots & u_{1,n} \\ u_{2,1} & u_{2,2} & \cdots & u_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ u_{i+1,1} & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ u_{c,1} & u_{c,2} & \cdots & u_{c,n} \end{bmatrix}$$

Clustering - Métodos por Particionamento

O processo de minimização deve obedecer as seguintes restrições:

Restrição 1

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j \in 1, \ldots, n.$$

garantindo que a soma das pertinências de um dado $\vec{x}_j$ a todos os grupos em C seja igual a 1, ou seja, cada coluna da matriz de partição deve possuir o valor 1 em **uma e somente uma** célula.

Restrição 2

$$\sum_{j=1}^{n} u_{ij} \geq 1, \forall i \in 1, \ldots, c.$$

tal que cada linha da matriz de partição deve possuir o valor 1 em **pelo menos** uma célula. Para garantir que todos os c grupos tenham, ao menos, um dado associado.

Clustering - Métodos por Particionamento

No processo de minimização de J_{CM}, tanto U_h quanto $\mathcal{C}$ devem ser atualizados:

Atualização de U_h

$$u_{ij}^{t+1} = \begin{cases} 1, & \text{se } i = \arg\min_{i=1}^{c} d(\vec{C}_i, \vec{x}_j) \\ 0, & \text{caso contrário.} \end{cases} \quad (2)$$

onde t é o contador de iterações do processo de otimização e u_{ij}^{t+1} é o valor da pertinência do dado j ao grupo i na iteração $t+1$. A atualização faz com que cada dado seja associado ao grupo cujo protótipo é o mais próximo a ele (possui a distância mínima) dentre todos os protótipos.

Atualização de $\mathcal{C}$

$$\vec{C}_i^{t+1} = \frac{\sum_{j=1}^{n} u_{ij} \vec{x}_j}{\sum_{j=1}^{n} u_{ij}} \quad (3)$$

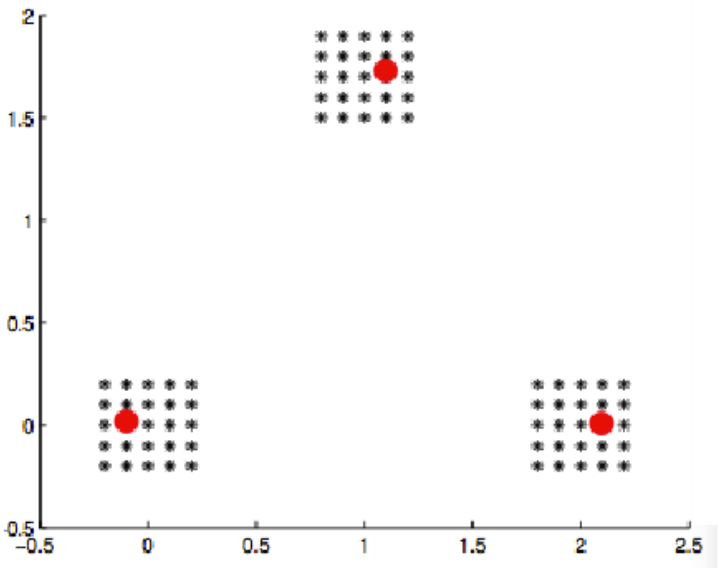
estabelece novos vetores protótipos para os grupos de acordo com a média de todos os vetores de dados associados a eles. O numerador soma, para cada grupo, os vetores de dados associados a eles. O denominador termina o processo de média.

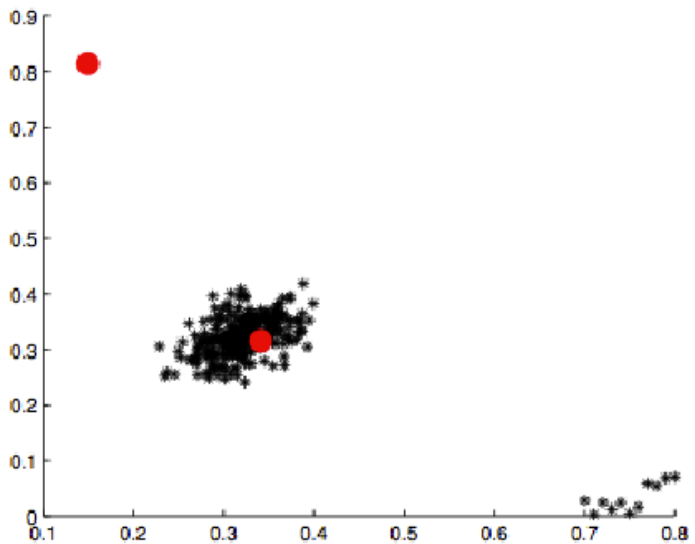
Clustering - Métodos por Particionamento

Algoritmo 1 *c-Means*

Determine a quantidade de partições c;
Determine um valor pequeno e positivo para um erro máximo, ϵ, permitido no processo;
Inicialize o conjunto de protótipos $\mathcal{C}$ aleatoriamente, escolhendo c vetores protótipos dentro do menor intervalo que contém todos os dados do conjunto; ou inicialize tais vetores escolhendo aleatoriamente c dados do conjunto de dados;
Inicialize o contador de iterações t como $t = 0$;
repita
  $t++$;
  Atualize U_h de acordo com a Equação (3);
  Atualize $\mathcal{C}$ de acordo com a Equação (4);
até que $\left\| \mathcal{C}^{(t)} - \mathcal{C}^{(t-1)} \right\| < \epsilon$

Clustering - Métodos por Particionamento





Nos exemplos, o primeiro caso foi bem resolvido pelo c-Means, já o segundo caso não foi bem resolvido.

Clustering - Métodos por Particionamento

c-Medoids

O c-Means é um algoritmo bastante sensível a ruído, e outliers podem distorcer a formação dos grupos. Uma forma de alterar o algoritmo para tentar diminuir essa sensibilidade é usar um dado (o mediano) para representar o grupo, ao invés de usar a média dos dados.

Assim o método de particionamento deixa de ser guiado por um princípio de minimização do erro de quantização e passa a ser guiado pela minimização das dissimilaridades entre os dados e o dado de referência.

O processo vai iterar até que o dado representativo esteja localizado "mais centralmente" (most centrally located) no seu cluster. Veja um algoritmo em Han & Kamber, página 406.

Clustering

Hierárquico ....

Clustering - Métodos Hierárquicos

Neste tipo de método de agrupamento, os dados são agrupados em "árvores". Os métodos podem ser **aglomerativos** ou **divisivos**, dependendo se a decomposição hierárquica é formada usando uma estratégia *bottom-up* (merge) ou *top-down* (*split*).

Esses algoritmos, em sua forma pura, sofrem do problema de não poderem executar ajustes uma vez que foi tomada uma decisão sobre juntar grupos ou dividir grupos. Isso pode levar à necessidade de alterar o método, mesclando-o com outras estratégias.

Clustering - Métodos Hierárquicos

Clustering Hierárquico Aglomerativo

A estratégia *bottom-up* inicia pela alocação de cada objeto em seu próprio cluster. Então, junções destes clusters (atômicos) são realizadas, formando clusters cada vez maiores, até que todos os objetos sejam alocados em um único cluster, ou alguma condição de parada seja satisfeita.

Clustering Hierárquico Divisivo

A estratégia *top-down* inicia com todos os objetos em um cluster. Então, divide o cluster em pedaços menores, até que cada objeto forme o seu próprio cluster, ou até que uma determinada condição de parada seja satisfeita (por exemplo, um número desejado de clusters, ou o diâmetro de cada cluster atingir um limiar).

Clustering - Métodos Hierárquicos

Abordagens usadas nos algoritmos hierárquicos

Na abordagem **single linkage**, cada cluster é representado por todos os objetos nele contidos, e a similaridade entre dois clusters é representada pela distância entre pares de dados (objetos) mais próximos e pertencentes a clusters diferentes.

A abordagem **complete linkage** usa a maior distância entre dois grupos. É o método do vizinho mais distante. A distância entre dois clusters é determinada de acordo com a maior distância entre um par de dados, sendo cada dado pertencente a um cluster distinto.

A abordagem **average linkage** usa a média entre as distâncias. Ou seja, é calculada a média das distâncias entre todos os pares de dados de dois grupos. Os pares de grupos que apresentarem a menor média são mais similares.

Centroid-linkage usa o vetor protótipo do cluster para o cálculo da similaridade entre clusters. A similaridade entre os clusters é definida com base na distância euclidiana entre os protótipos dos clusters.

Clustering - Métodos Hierárquicos

AGNES (AGglomerative NESting) X DIANA (DIvisive ANAlysis)

AGNES: Inicialmente coloca cada objeto em seu próprio cluster, e depois junta os clusters, passo a passo, de acordo com algum critério (por exemplo, usando a abordagem *single-linkage*, cluster C1 e C2 se juntam se um objeto em C1 e um objeto em C2 possuem a distância Euclidiana mínima entre quaisquer dois objetos de clusters diferentes.) O processo de união (*merge*) continua até que só exista um cluster.

DIANA: Todos os objetos são usados para formar um cluster inicial. Para a divisão, o dado menos semelhante a todos os outros é selecionado para conduzir a formação de um novo cluster. Então, são buscados dentro do cluster original, os elementos que são mais semelhantes (de acordo com uma métrica de similaridade) ao novo cluster do que ao cluster original. Esses dados são transladados para o novo grupo.

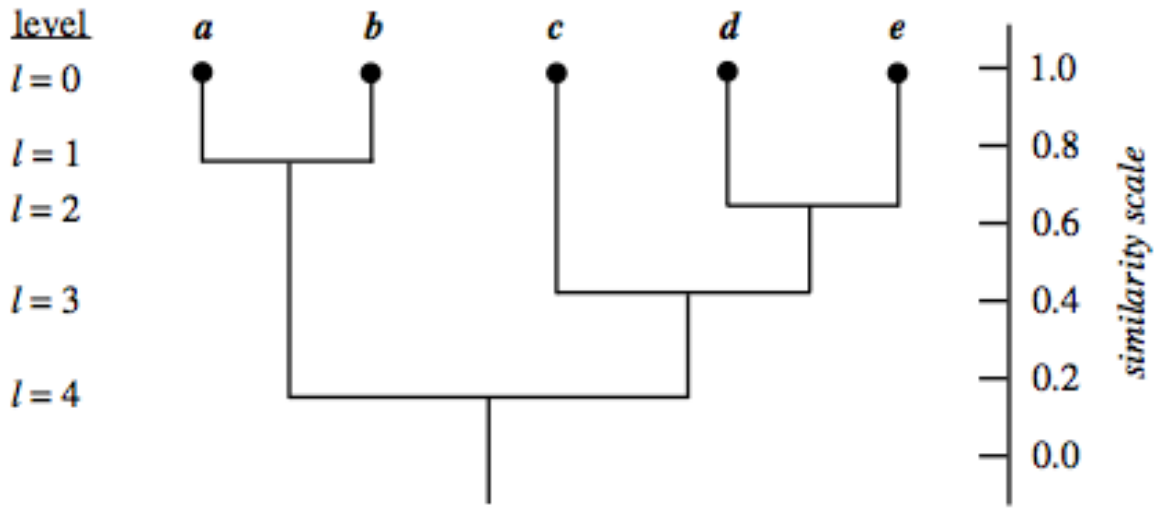
Método Aglomerativo $\rightarrow$ AGNES (na folha impressa)

Método Divisivo $\rightarrow$ DIANA (na folha impressa)

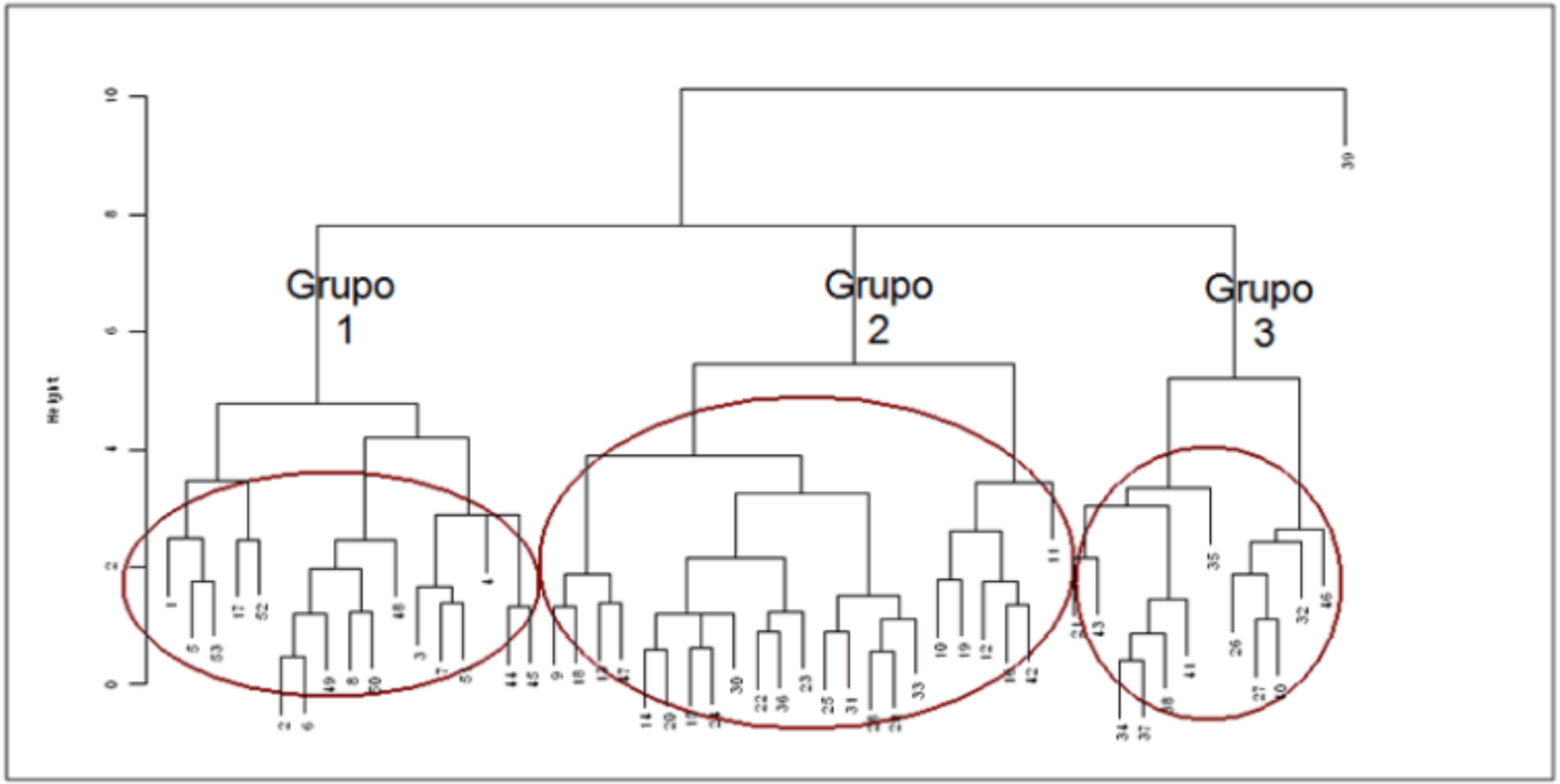
Clustering - Métodos Hierárquicos

Visualização

Os clusters criados por algoritmos hierárquicos podem ser visualizados por meio de um dendograma.



Clustering - Métodos Hierárquicos



http://www2.inecc.gob.mx/publicaciones/libros/496/cap3.html

Clustering - Métodos Hierárquicos

BIRCH: Balanced Iterative Reducing and Clustering Using Hierarchies

Projetado para trabalhar com conjuntos de dados numéricos grandes. Integra a estratégia de clustering hierárquico (na fase inicial – microclustering) e outros métodos de clustering (na fase final – macroclustering). Essa estratégia supera duas dificuldades do método de clustering aglomerativo: (1) escalabilidade e (2) inabilidade de desfazer o que foi feito em passos anteriores.

Conceitos:

- *clustering feature* (CF): é um vetor tridimensional que resume informação sobre os objetos de um cluster e é definido como $CF =< n, LS, SS >$, onde n é o número de pontos em um cluster, LS é a soma dos n pontos e SS é a soma quadrada dos n pontos.
- *clustering feature tree* (CF tree): é uma árvore balanceada que estoca os CFs para uma clusterização hierárquica.

Clustering - Métodos Hierárquicos

BIRCH - Clustering Features

Um CF é um resumo de estatísticas para um dado cluster, e tais características são aditivas. Então, se temos dois clusters disjuntos, C_1 e C_2, que possuem os CF_1 e CF_2, respectivamente, um novo cluster composto pela junção de C_1 e C_2, será simplemente $CF_1 + CF_2$.

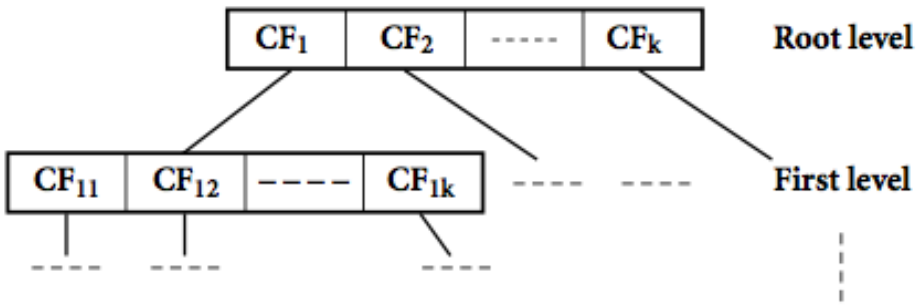
CFs são suficientes para calcular todas as medidas que são necessárias para as tomadas de decisões realizadas no algoritmo BIRCH. Assim, não é necessário trabalhar com os dados todo o tempo.

Clustering - Métodos Hierárquicos

BIRCH - CF Tree

Os nós não folha estocam somas de CFs de seus filhos, e portanto, sumarizam a informação de clusterização contida nos seus filhos. A árvore conta também com dois parâmetros, os quais influenciam no tamanho resultante da árvore:

- Fator de ramificação (B): especifica o número máximo de filhos por nó não folha;
- limiar (T): especifica o diâmetro máximo de subclusters estocados nos nós folhas da árvore.



Clustering - Métodos Hierárquicos

BIRCH

O algoritmo BIRCH tenta produzir os melhores clusters com os recursos disponíveis. Para isso, o algoritmo aplica uma técnica multifase com uma leitura única do conjunto de dados, e uma ou mais leituras adicionais se for necessário para melhorar a qualidade do resultado.

- Fase 1: o algoritmo lê o conjunto de dados e constrói uma CF tree inicial, a qual pode ser vista como uma compressão multinível dos dados que tenta preservar características da estrutura de cluster existente nos dados.
- Fase 2: o algoritmo aplica um segundo algoritmo para clusterizar os nós folhas da CF tree, o qual remove clusters esparsos como outliers e agrupa clusters densos dentro de outros ainda maiores.

Clustering - Métodos Hierárquicos

BIRCH - Fase 1

Na fase 1, a CF tree é construída incrementalmente, conforme os dados são inseridos nela. Um objeto é inserido na folha mais próxima. Se o diâmetro do subcluster estocado no nó folha depois da inserção do atual dado é maior do que um valor limiar, então um novo nó folha deve ser criado. Depois da inserção de um novo objeto, a informação sobre ele deve ser propagada até a raíz da árvore.

Procedimento de inserção

- A partir da raíz, encontre a folha apropriada para inserção: siga o caminho CF *mais próximo* usando uma métrica de similaridade (por exemplo, a distância euclidiana do dado para o centróide do cluster de uma "chave" do nó);
- Modifique o nó folha encontrado: se o nó folha mais próximo não pode receber o dado (o cluster deste nó já alcançou o diâmetro máximo), crie um novo nó folha. Se não houver espaço para o novo nó, quebre o nó pai (como em uma árvore B);
- atualize os CFs como resultado •apenas da inserção do dado ou como resultado da inserção e da quebra de um nó.

Clustering - Métodos Hierárquicos

BIRCH - medidas

Dado n pontos x_i (dados) num espaço d-dimensional, ou pontos em um cluster, define-se o centróide x_0, o raio R, e o diametro D do cluster conforme especificado abaixo:

$$x_0 = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4}$$

$$R = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - x_0)^2} \tag{5}$$

$$D = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2} \tag{6}$$

R é a distância média dos objetos ao centróide do cluster e D é a distância par a par média dentro de um cluster. Ambos medem o espalhamento (ou a concentração) dos dados ao redor do centróide do cluster.

Clustering - Métodos Hierárquicos

uma fase 1b ...

Escolha um T (limiar para o diâmetro) maior e construa uma nova árvore reorganizando os CF nas folhas. Com T maior um CF pode receber mais dados e grupos poderão se juntar. A árvore ficará menor.

BIRCH - Fase 2

Na fase 2, considere as entradas CF dos nós folha. Use os centróides como protótipos dos clusters. Execute uma clusterização tradicional (baseadas nos protótipos, e não em medidas que resumem a informação de um cluster) - ou seja, considere os representantes dos CFs e não os dados do conjunto original.

uma fase 2b ...

Considere o conjunto de dados original (mais uma vez) e use os clusters encontrados na fase 2 como sementes. Redistribua os dados às sementes mais próximas. Remova os outliers (sementes que não recebem dados). Com isso você obtem a informação de pertinência dos dados originais aos clusters.

Clustering - Métodos Hierárquicos

BIRCH - paper

BIRCH: An Efficient Data Clustering Method for Very Large Databases. Tian Zhang, Radgu Ramakrishnan, Miron Livny. SIGMOD, 1996. Conferência do ACM Special Interest Group on Management of Data.

Clustering

Densidade ....

Clustering - Métodos Baseados em Densidade

Os métodos de agrupamento baseados em densidade tentam suprir a necessidade de métodos capazes de descobrir grupos com formas arbitrárias. Nestes algoritmos, a ideia de grupos é baseada na existência de regiões densas de dados, separadas por regiões com baixa densidade de dados.

Alguns exemplos de algoritmos desta classe são:

- **DBSCAN**: Density-Based Spatial Clustering of Applications with Noise;
- OPTICS: Ordering Points to Identify the Clustering Structure;
- DENCLUE: Density-based Clustering;

Clustering - Métodos Baseados em Densidade

DBSCAN

Princípio: o processo executado no algoritmo "encontra" regiões com densidade suficientemente alta para descobrir os clusters, considerando um conjunto de dados "com ruído". Neste algoritmo, um cluster é definido como o um conjunto máximo de *density-connected points*.

Um cluster baseado em densidade é um conjunto de objetos conectados "por densidade" que é máximo com respeito à densidade alcançável. Todo objeto não contido em um cluster é considerado ruído.

Clustering - Métodos Baseados em Densidade

DBSCAN - definições

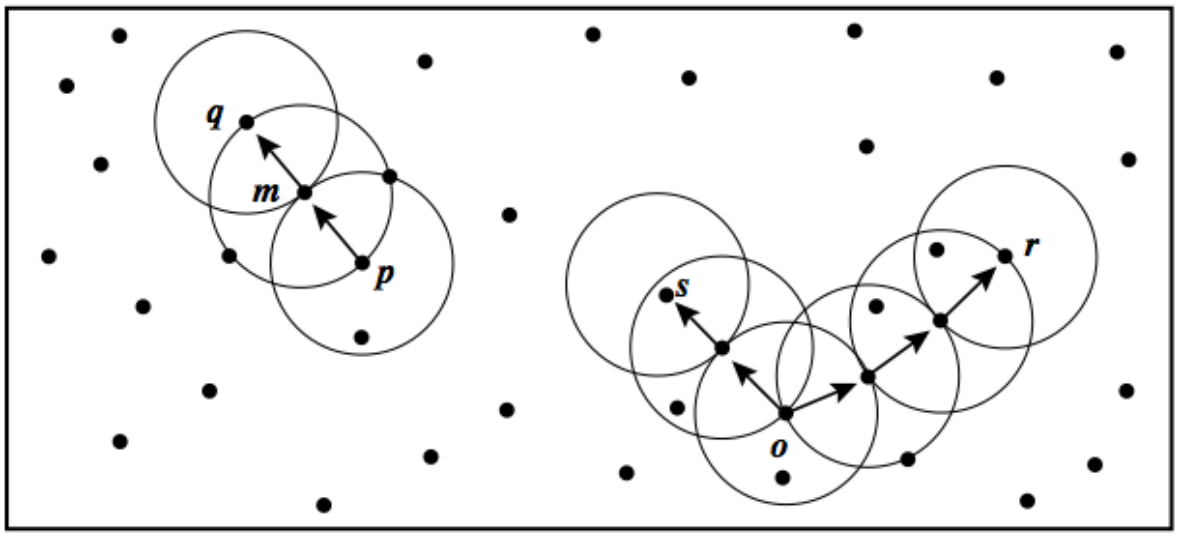
- a vizinhança dentro de um raio ϵ de um dado objeto é chamada de ϵ-*vizinhança* do objeto;
- se a ϵ-*vizinhança* de um objeto contém pelo menos um número mínimo de pontos, *MinPts* (ou de objetos), então o objeto é chamado de **objeto núcleo**; cc. ele é um **objeto de borda**.
- dado um conjunto de objetos, D, um objeto p é **diretamente alcançavel "por densidade"** a partir de um objeto q, se p está dentro da ϵ- *vizinhança* de q e q é um objeto núcleo;
- um objeto p é **alcançável por densidade** a partir do objeto q considerando ϵ e *MinPts* em um conjunto de objetos D, se existe uma cadeia de objetos $p_1, ..., p_n$ onde $p_1 = q$ e $p_n = p$ tal que p_{i+1} é diretamente alcançável de p_i considerando ϵ e *MinPts*, para $1 \leq i \leq n$, $p_i \in D$.
- um objeto p é **conectado "por densidade"** ao objeto q considerando ϵ e *MinPts* em um conjunto de objetos D, se existe um objeto $o \in D$ tal que tanto p quanto q são alcançáveis "por densidade" a partir de o considerando ϵ e *MinPts*.

Clustering - Métodos Baseados em Densidade

DBSCAN

Analise a figura abaixo, onde $MinPts = 3$ e ϵ é representado pelo raio dos círculos. Considerando os objetos rotulados:

- quais objetos são objetos núcleos?
- quais objetos são diretamente alcancáveis por densidade? quais não são?
- quais objetos são alcançáveis por densidade a partir de quais objetos? quais não são?
- quais objetos sao conectados por densidade?



Clustering - Métodos Baseados em Densidade

DBSCAN

- **m**, **p**, **o** e **r** são objetos núcleos;
- **q** é diretamente alcançável por densidade a partir de **m**. **m** é diretamente alcançável por densidade a partir de **p** e vice-versa.
- **q** é (indiretamente) alcançável por densidade a partir de **p** porque **q** é diretamente alcancável por densidade a partir de **m** e **m** é diretamente alcançavel por densidade a partir de **p**. Contudo, **p** não é diretamente alcançavel por densidade a partir de **q** porque **q** não é um objeto núcleo. Similarmente, **r** e **s** são alcancáveis por densidade a partir de **o**, e **o** é alcançavel por densidade a partir de **r**.
- **o**, **r**, e **s** são todos conectados por densidade.

Clustering - Métodos Baseados em Densidade

DBSCAN - procedimento

- É preciso definir ϵ e *MinPts*.
- DBSCAN procura pelos clusters checando a vizinhança de cada ponto no conjunto de dados.
- Se a ϵ-vizinhança contém mais do que *MinPts*, um novo cluster com **p** como objeto núcleo é criado.
- DBSCAN iterativamente encontra os objetos diretamente alcançáveis por densidade a partir destes objetos núcleos, o que deve ocasionar o merge de alguns clusters.
- O processo termina quando nenhum novo ponto pode ser adicionado ao algum cluster.

Algoritmo $\rightarrow$ na folha impressa

Clustering - Métodos Baseados em Densidade

BDSCAN - Altamente sensível aos parâmetros iniciais

epsilon	*MinPnt*	Resultado
Alto	Alto	Poucos clusters, grandes e densos.
Baixo	Alto	Mais clusters, pequenos e densos
Alto	Baixo	Menos clusters, grandes e pouco densos
Baixo	Baixo	Muitos clusters, pequenos e pouco densos

Clustering - Métodos Baseados em Densidade

BDSCAN - paper

A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu. Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96) 1996.

Clustering

Grid ....

Clustering - Métodos Baseados em Grid

Esta abordagem usa uma estrutura de dados em grade (de multiresolução) para encontrar os clusters. Ela quantiza o espaço dos dados em um número finito de células que forma uma estrutura em grid na qual todas as operações para clustering são executadas.

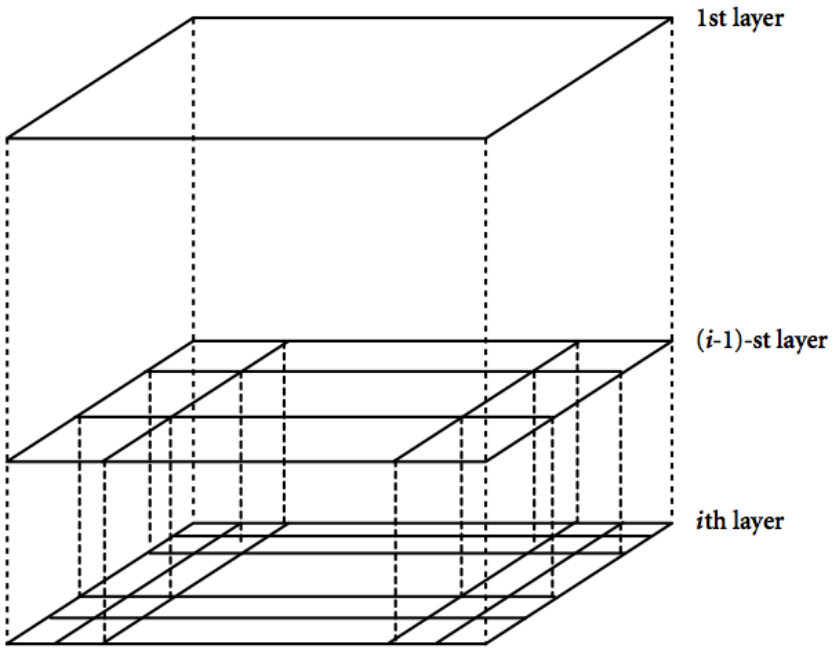
Exemplos

- **STING**: STatistical INformation Grid;
- WaveCluster: Clustering Using Wavelet Transformation;
- CLIQUE: Clustering In QUEst.

Clustering - Métodos Baseados em Grid

STING

Nesta técnica, a área espacial dos dados (o espaço dos dados) é dividido em células retangulares. Diferentes níveis de células correspondem a diferentes níveis de resolução, e isto forma uma ideia de estrutura hierárquica: cada célula de nível mais alto é particionada para formar uma quantidade de células do próximo nível. Então, informações estatísticas referente a atributos das células são pré-computadas e armazenadas. Estes parâmetros são usados em um processo de consulta.



Clustering - Métodos Baseados em Grid

STING

Parâmetros estatísticos de células de nível mais alto podem ser calculados a partir daqueles já computados para as células de nível mais baixo. Estes parâmetros incluem:

- parâmetro independente de atributo: número de dados nas células;
- parâmetros dependentes de atributo: média, desvio padrão, mínimo, máximo, calculados para cada atributo do dado;
- tipo de distribuição seguida pelos valores dos atributos: normal, uniforme, exponencial, ..., nenhuma (ou desconhecida) – também calculado por atributo.

STING X agrupamento

O problema de agrupamento é resolvido por meio de consultas que são respondidas a partir de buscas nas informações armazenadas na estrutura de grid do STING. Por exemplo: recupere regiões em que a densidade de pontos é maior que x e que o atributo y tenha média z.

A busca se dá sobre as informações estatísticas, iniciando nas células de um nível arbitrado. Células interessantes são marcadas e então, na próxima iteração, suas filhas são examinadas.

Clustering - Métodos Baseados em Grid

STING - paper

STING: A Statistical Information Grid Approach to Spatial Data Mining. Wei Wang, Jiong Yang, Richard Muntz. Proceedings of the 23rr VLDB Conference 1997.

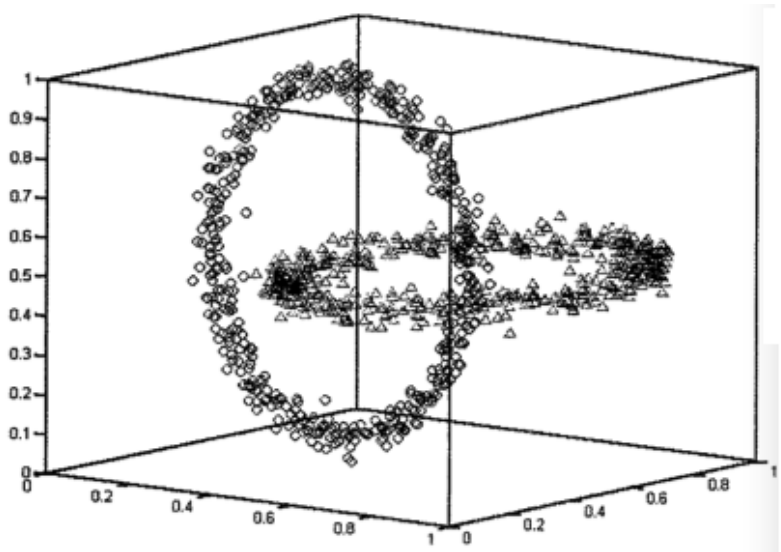
Clustering

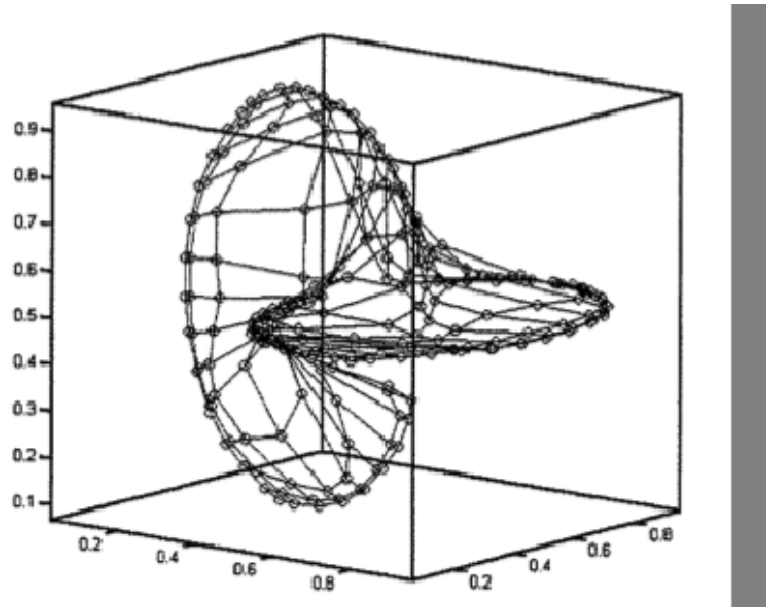
Modelo ....

Clustering - Métodos baseados em Modelo

Self Organizing Maps (SOM) - Mapas Auto Organizáveis

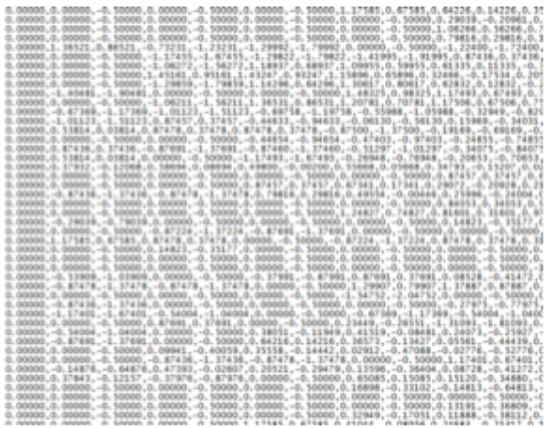
O SOM foi inspirado no modo pelo qual informações sensoriais são mapeadas no córtex cerebral. SOM é um algoritmo não supervisionado que aproxima a densidade de probabilidade dos dados de entrada ao mesmo tempo em que reduz a dimensionalidade, tentando preservar ao máximo as relações topológicas entre os dados.





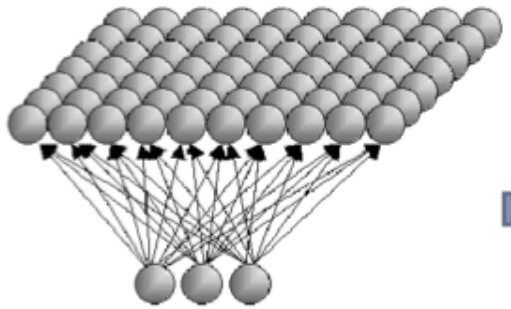
Clustering - Métodos baseados em Modelo

Motivando o SOM...

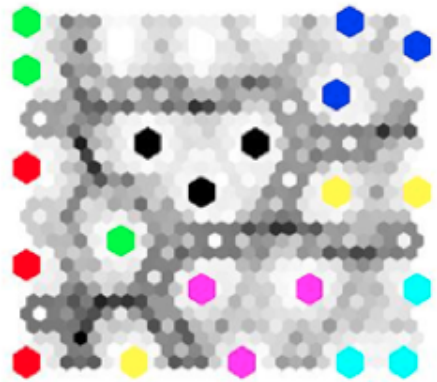


Conjunto de dados
(n-dimensional)

Som bi-dimensional



Pós-
processamento

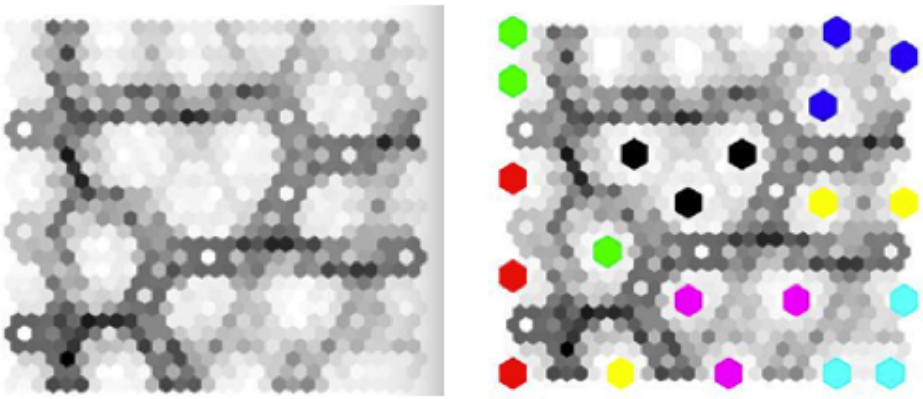


Matriz U (rotulada)

Quantização do espaço
dos dados.

Clustering - Métodos baseados em Modelo

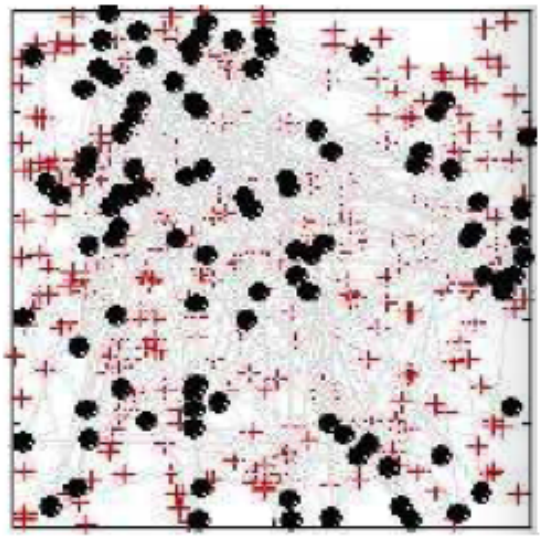
Objetivos do SOM ...

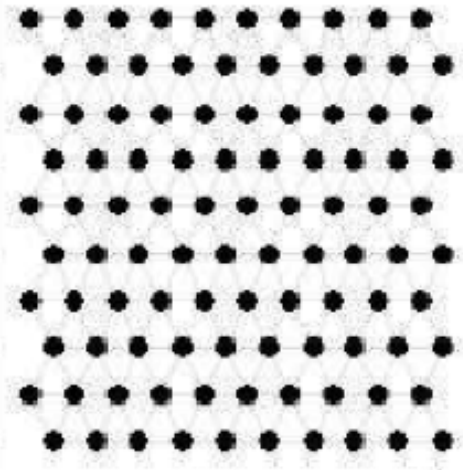


- **Aproximação do espaço de entrada**: Um dos objetivos de um SOM é representar um conjunto grande de vetores de entrada localizados em um espaço de alta dimensão, por meio de um conjunto menor de vetores localizados em um espaço de dimensão mais baixa. Ou seja, realizar a **quantização do espaço** e a **redução de dimensão**.
- **Visualização do conjunto de dados**: Visualizar como se dão as relações espaciais entre os dados. Quando é este o objetivo pode-se desconsiderar a necessidade de quantização do espaço.

Clustering - Métodos baseados em Modelo

Conceituando os espaços no contexto do SOM ...

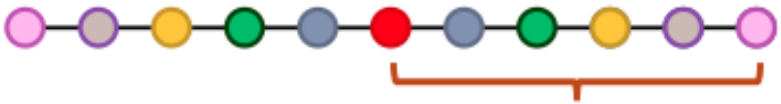




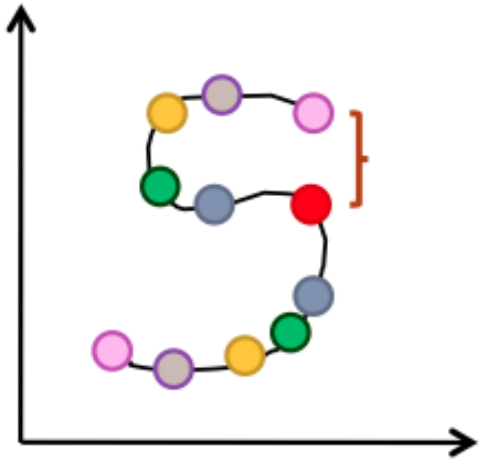
- **Espaço de entrada** ou espaço dos dados: espaço vetorial onde os nerônios da rede SOM residem. Distâncias vetoriais são usadas neste espaço para analisar a similaridade entre vetores (neurônios e/ou dados).
- **Espaço de saída:** espaço matricial onde é organizada a topologia da rede SOM. Distâncias matriciais são usadas neste espaço para analisar as relações de vizinhança entre os vetores (neurônios).

Clustering - Métodos baseados em Modelo

Conceituando vizinhança e topologia no SOM ...

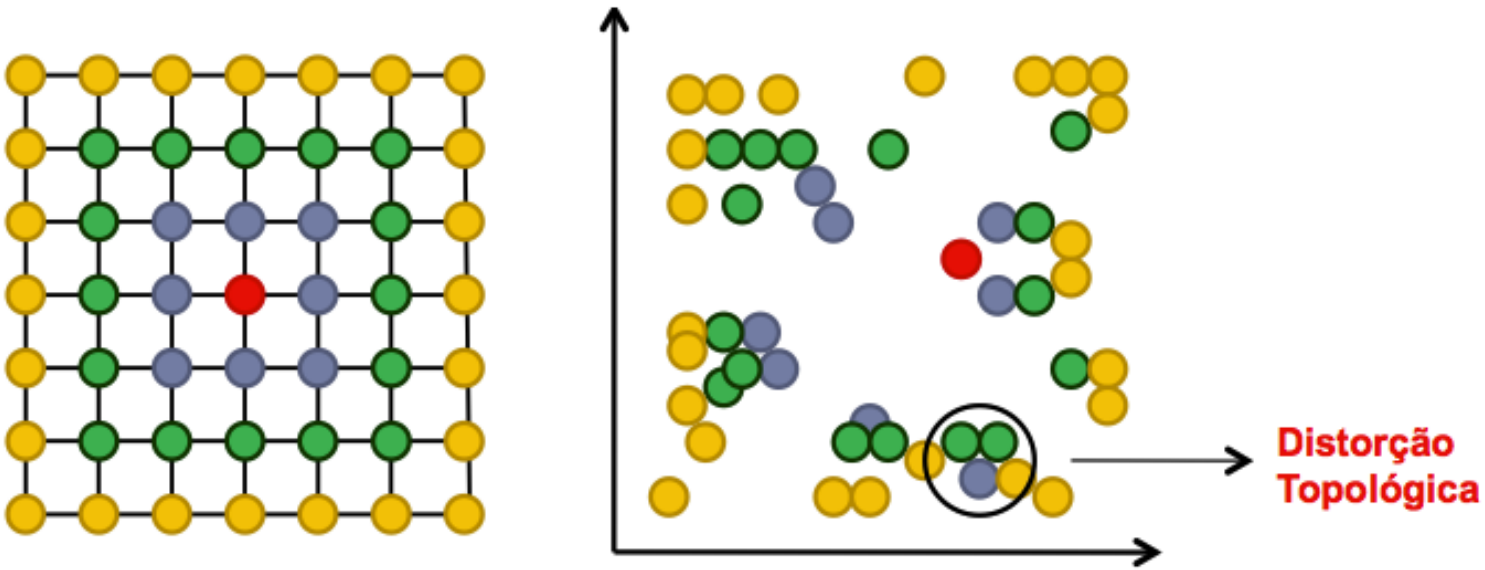


Raio de vizinhança = 5



Clustering - Métodos baseados em Modelo

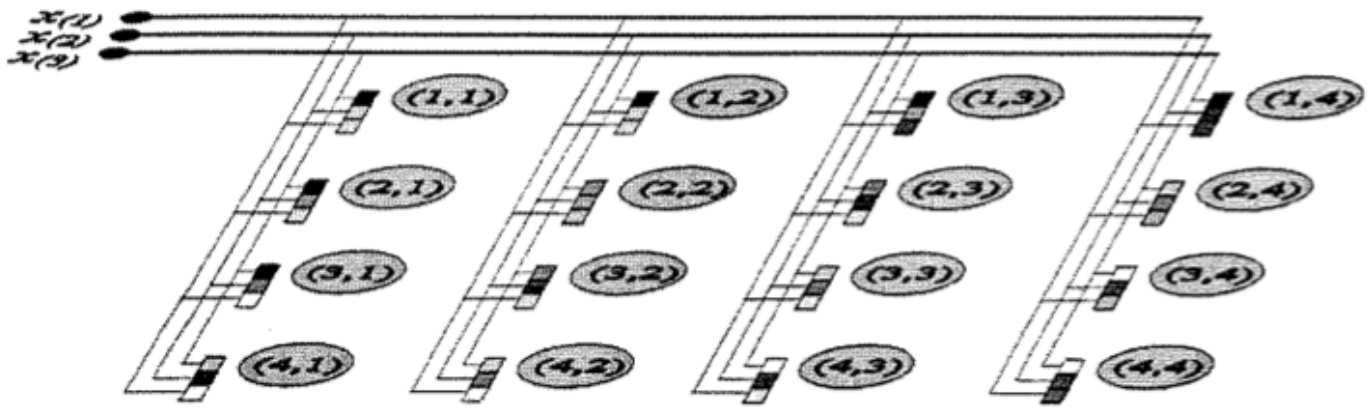
Conceituando vizinhança e topologia no SOM ...



Clustering - Métodos baseados em Modelo

Definindo a arquitetura da rede SOM ...

- número de neurônios na camada de entrada
- número de neurônios na cadama de saída (tamanho do mapa)
- tipo de vizinhança topológica (dimensão e lattice)
- função de vizinhança (...)



- 3 neurônios na camada de entrada (espaço vetorial tridimensional)
- 16 neurônios na camada de saída
- mapa bidimensional (espaço matricial bidimensional)
- lattice retangular

Clustering - Métodos baseados em Modelo

Algoritmo de treinamento

* Notação usada na Fausett.

- Passo 0:
  - Determine a arquitetura da rede neural SOM
  - Inicialize os pesos w_{ij} % posicionar os neurônios no espaço vetorial
  - Determine os parâmetros da taxa de aprendizado (valor inicial e função de atualização)
- Passo 1: Enquanto condição de parada é falsa, execute os passos 2-8
  - Passo 2: Para cada vetor de entrada x, execute os passos 3-5
    - Passo 3: Para cada j, compute: $D(j) = \sum_i (w_{ij} - x_i)$
    - Passo 4: Encontre o J tal que $D(J)$ seja mínimo
    - Passo 5: Para todas as unidades j dentro de uma vizinhança específica de J, e para todo i: $w_{ij}(new) = w_{ij}(old) + \alpha[x_i - w_{ij}(old)]$
  - Passo 6: Altere a taxa de aprendizado (se for o caso)
  - Passo 7: Reduza o raio de vizinhança (se for o caso)
  - Passo 8: Teste a condição de parada

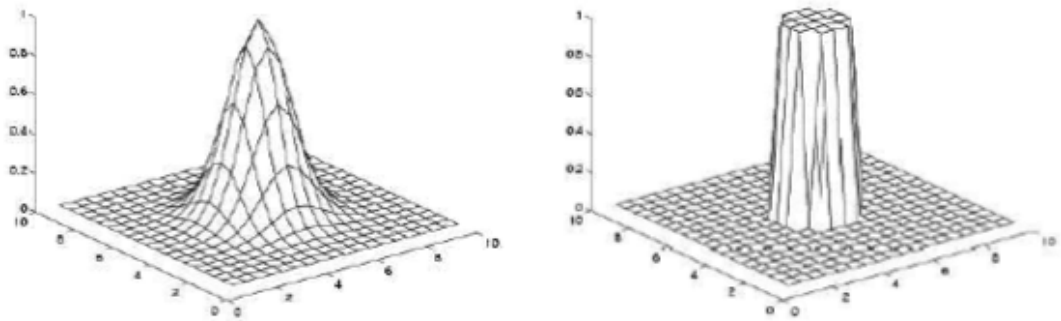
Clustering - Métodos baseados em Modelo

Inicialização de pesos

- **Aleatória**: posiciona os neurônios do mapa de forma aleatória dentro do espaço dos dados.
- **Linear**: usa os componentes principais da matriz de autocorrelação do conjunto de dados X. As posições dos neurônios são determinadas de forma a distribuir-se na direção dos espaços dos autovetores correspondentes aos maiores autovalores encontrados. Os neurônios se distribuem nas direções de maior variância dos dados.
- **Usando conhecimento** *a priori* : usa algum conhecimento sobre os dados para posicionar os neurônios em locais adequados dentro do espaço dos dados.

Clustering - Métodos baseados em Modelo

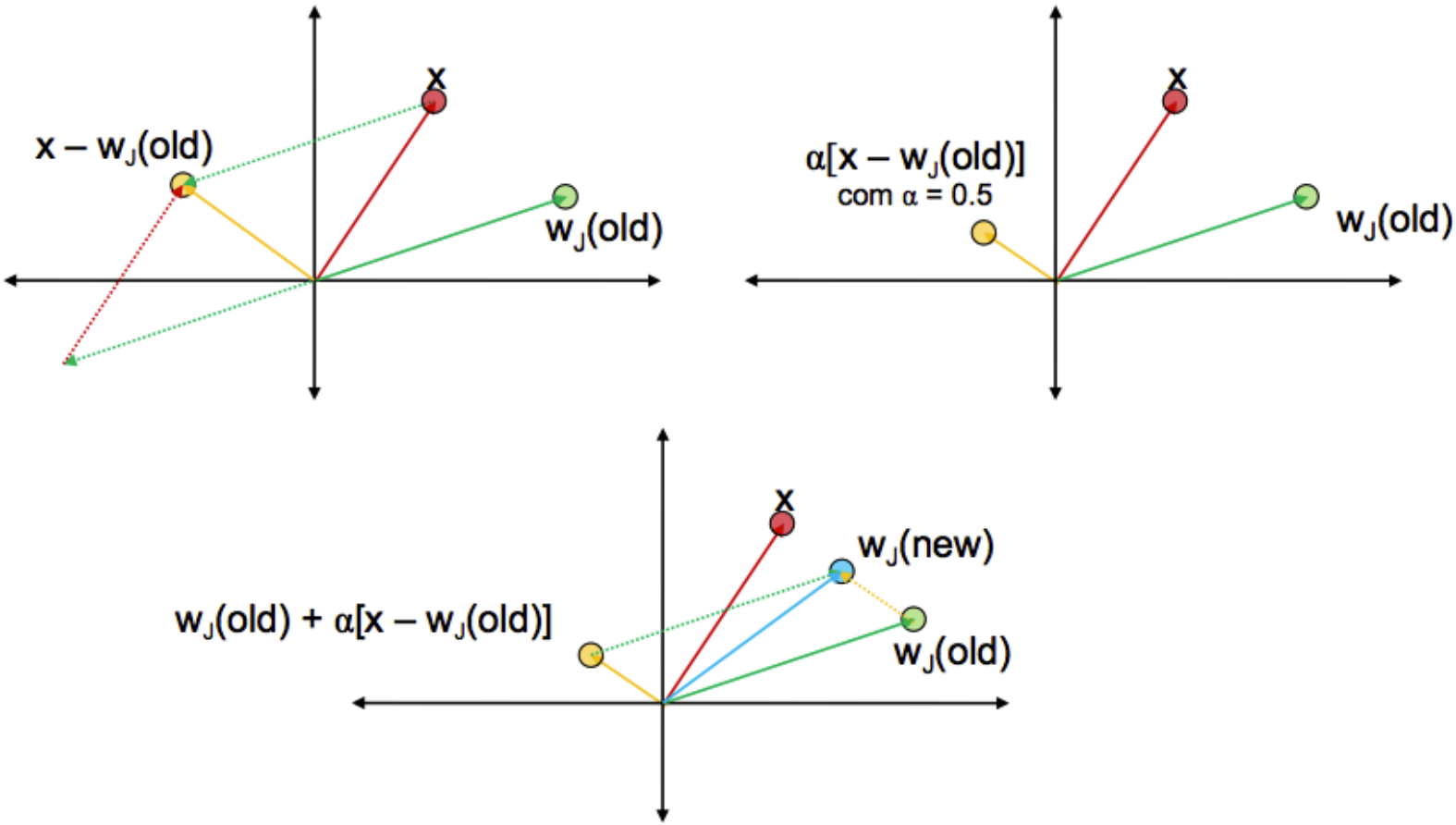
Analisando funções de vizinhança ... ...



$w_i(new) = w_{ij}(old) + g(j)\alpha[x_i - w_{ij}(old)]$ $w_i(new) = w_{ij}(old) + b(j)\alpha[x_i - w_{ij}(old)]$

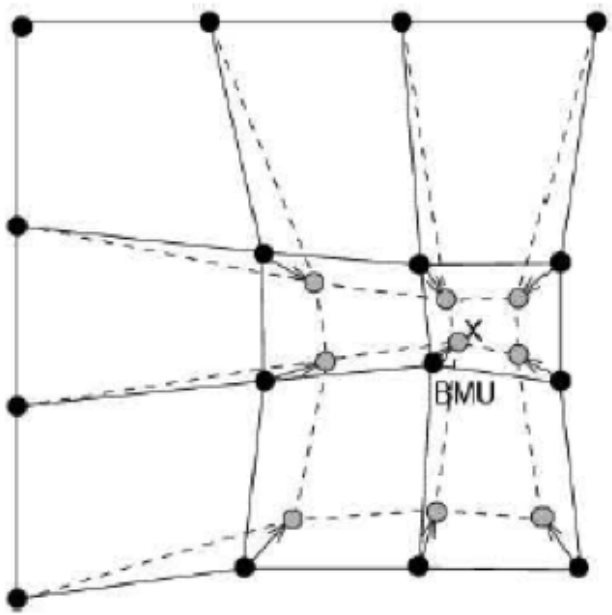
Clustering - Métodos baseados em Modelo

Estudando a taxa de aprendizado ... ...



Clustering - Métodos baseados em Modelo

Analisando o ajuste de pesos ... ...



$w_i(new) = w_{ij}(old) + b(j)\alpha[x_i - w_{ij}(old)]$

* j varia na vizinhança do neurônio BMU

Clustering - Métodos baseados em Modelo

Outras decisões e conceitos ....

- Número de épocas
  - fase de ordenação
  - fase de sintonização (ajuste fino)
- Função de atualização da taxa de aprendizado
- Função de atualização do raio de vizinhança
- Algoritmo de treinamento em lote (batch)
- BMU - Best Matching Unit

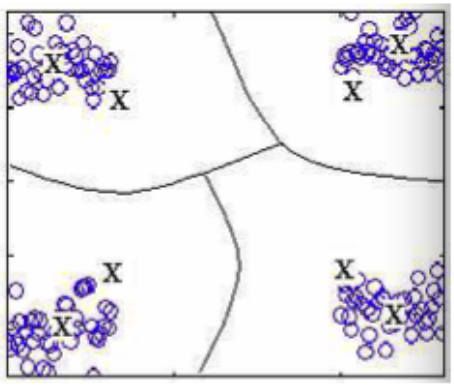
Clustering - Métodos baseados em Modelo

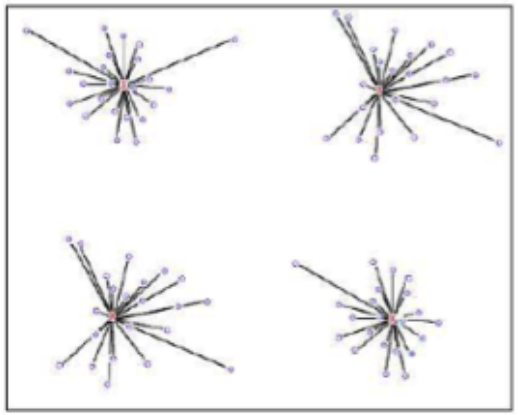
Qualidade do SOM - Erro de quantização

O número de neurônios no mapa deve ser menor (bem menor) que o número de dados no conjunto de dados estudado para que se tenha um alto grau de quantização. Mas na quantização, existe perda de informação e um erro é produzido:

$$E_q = \frac{1}{N} \sum_{i=1}^{N} ||x_i - w_{bi}||$$

onde N é o número de dados sob análise, b é o índice do BMU e $||.||$ é a distância entre os vetores (dado e BMU).





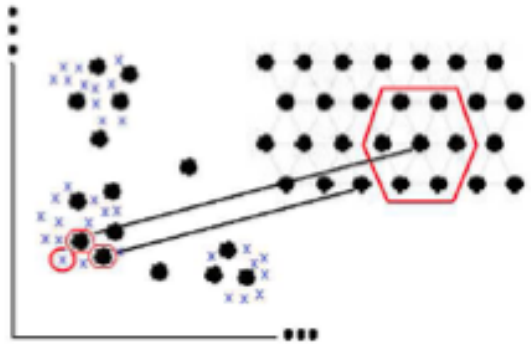
Clustering - Métodos baseados em Modelo

Qualidade do SOM - Erro Topológico

Ao reduzir a dimensionalidade do dos dados, um erro topológico é inserido no mapeamento.

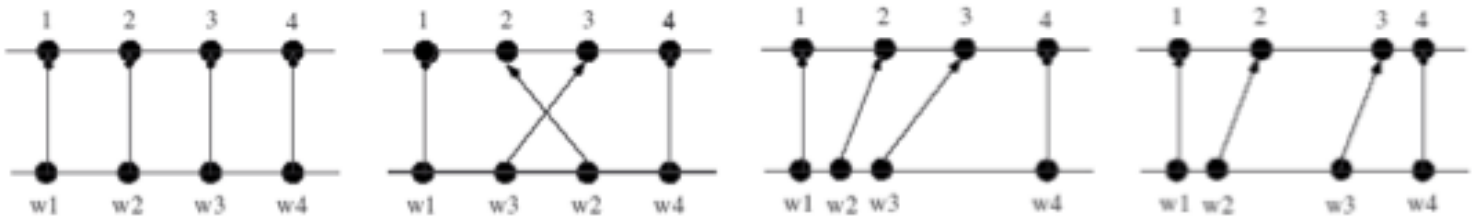
$$E_t = \frac{1}{N} \sum_{i=1}^{N} u(x_i)$$

onde $u(x_i) = 0$ se o primeiro e o segundo BMUs para o dado x_i forem adjacentes no espaço de saída; $u(x_i) = 1$ caso contrário.



Clustering - Métodos baseados em Modelo

Outras distorções topológicas ...

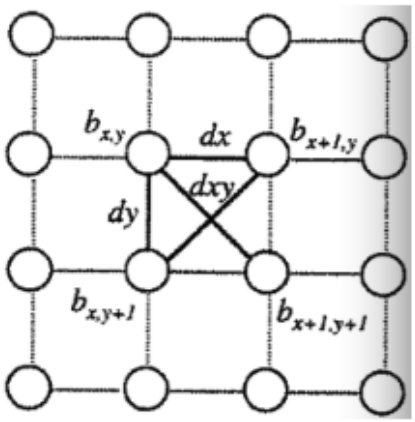


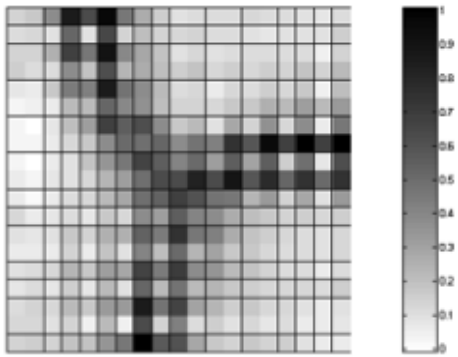
Clustering - Métodos baseados em Modelo

U_ matrix

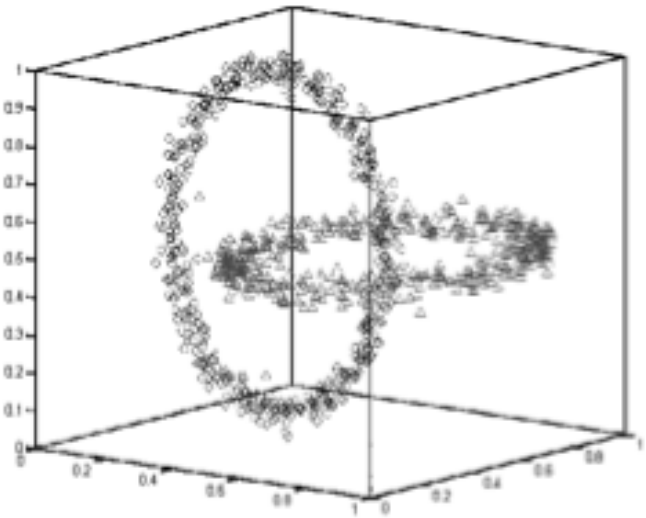
Método de visualização de um SOM treinado desenvolvido com o objetivo de permitir a detecção visual das relações topológicas dos neurônios. A ideia básica é usar a mesma métrica que foi usada durante o treinamento para calcular distâncias entre neurônios adjacentes.

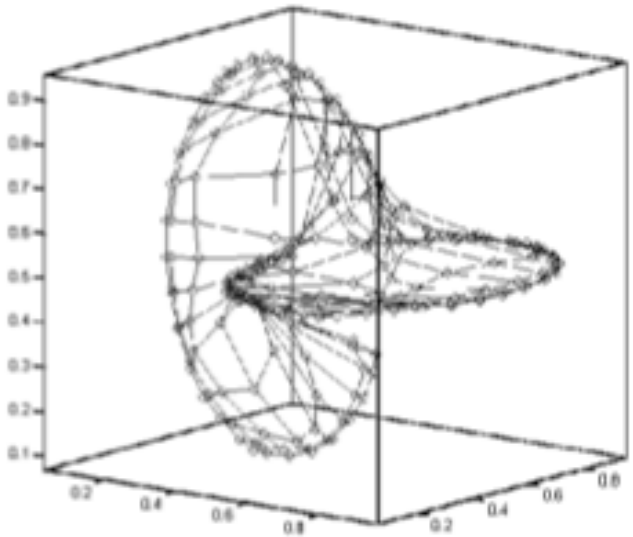
O resultado é uma imagem $f(x, y)$, na qual as coordenadas de cada pixel (x, y) são derivadas das coordenadas dos neurônios no grid do mapa, e a intensidade de cada pixel na imagem corresponde à uma distância calculada.

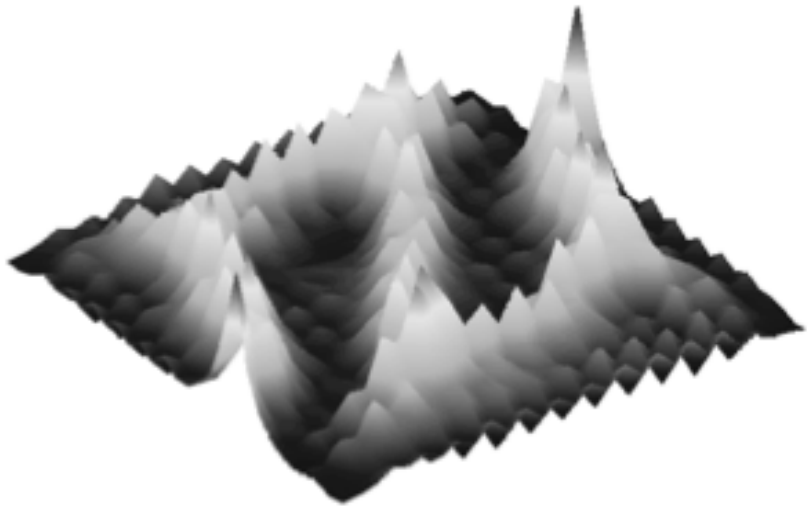




Clustering - Métodos baseados em Modelo







Clustering

Avaliação ....

Clustering - Comparativo Resumido

Nome	Tipo de dados	Geometria	Parâmetros de entrada
		Particionais	
k-means	Numérico	Formas não convexas	Número de grupos
k-modes	Categórico	Formas não convexas	Número de grupos
PAM	Numérico	Formas não convexas	Número de grupos
CLARA	Numérico	Formas não convexas	Número de grupos
CLARANS	Numérico	Formas não convexas	Número de grupos e número máximo de vizinhos
FCM	Numérico	Formas não convexas	Número de grupos

Clustering - Comparativo Resumido

Nome	Tipo de dados	Geometria	Parâmetros de entrada
		Hierárquicos	
BIRCH	Numérico	Formas não convexas	Raio do grupo e fator de ramificação
CURE	Numérico	Formas arbitrárias	Número de grupos e número de grupos representativos
ROCK	Categórico	Formas arbitrárias	Número de grupos
		Baseado em Densidade	
DBSCAN	Numérico	Formas arbitrárias	Raio do grupo e número mínimo de objetos
DENCLUE	Numérico	Formas arbitrárias	Raio do grupo e número mínimo de objetos

Clustering - Comparativo Resumido

Nome	**Tipo de dados**	**Geometria**	**Parâmetros de entrada**
		Baseado em Grid	
WaveCluster	Dados espaciais	Formas arbitrárias	Wavelets, número de células por dimensão, número de aplicações da transformada
STING	Dados espaciais	Formas arbitrárias	Número de objetos na célula, *fator de divisão da célula*

Clustering - Comparativo Completo

On Clustering Validation Techniques Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis Journal of Intelligent Information Systems, 17:2/3, 107-145 2001

Clustering - Avaliação

Validação de clusters

O problema de avaliar os resultados de um algoritmo de agrupamento é conhecido como **validação de grupos** (ou *cluster validity*). Podemos considerar três abordagens para isso:

- critérios externos: avaliação dos resultados com base em uma estrutura pré-especificada, a qual é imposta ao conjunto de dados e reflete uma intuição sobre a estrutura de grupos;
- critérios internos: avaliação dos resultados em termos quatitativos que envolvem os próprios vetores de dados.
- critérios relativos: avaliação dos resultados por comparação com outros esquemas de agrupamento, resultantes de outras execuções do mesmo algoritmo mas com parâmetros valorados de formas diferentes. A partir disso, aplicando:
  - compacidade: os membros de cada grupo deveriam ser tão próximos entre si quanto possível. Uma medida comum para a compacidade é a variância, a qual deve ser minimizada.
  - separação: os grupos demver ser largamente espaçados entre si. As medidas comums para isso são: single linkage, complete linkage, average linkage e centroid-linkage.

Clustering - Avaliação

Índices

- Critérios externos
  - usando Monte Carlo
  - **comparação de estrutura de grupos com partições de dados (não é válido para hierarquia de grupos)**
  - comparação de matriz de proximidade com partições de dados
- Critérios internos
  - validação hierárquica de esquema de grupos
  - validação de um único esquema de grupos
- Critérios relativos
  - agrupamento crisp (*modified Humbert* Γ *statistic*; **Dunn and Dunn-like indices**; Davies-Bouldin (DB) Index, RMSSDT, SPR, RS, CD);
  - agrupamento fuzzy (*partition coefficient*; *partition entropy coefficient*; *Xie-Beni index*; *Fukuyama-Sugeno index*; *fuzzy hyper volume*; *average partition density*)

Clustering - Avaliação

Comparação de estrutura de grupos com partições de dados (não é válido para hierarquia de grupos

Considere $C = C_1, ..., C_m$ uma estrutura de grupos de um conjunto de dados X; e $P = P_1, ..., P_s$ uma partição (definida usando algum tipo de conhecimento *a priori*) sobre os dados. Nós nos referimos a pares de pontos (x_v, x_u) usando os seguintes termos:

- **SS**: se os pontos pertencem ao mesmo grupo da estrutura de grupos C e ao mesmo grupo da partição P;
- **SD**: se os pontos pertencem ao mesmo grupo de C e a diferentes grupos em P;
- **DS**: se os pontos pertencem a diferentes grupos de C e para o mesmo grupo de P;
- **DD**: se os pontos pertencem a diferentes grupos de C e para diferentes grupos de P.

Clustering - Avaliação

Comparação de estrutura de grupos com partições de dados (não é válido para hierarquia de grupos

Assumindo:

- a = SS; b = SD; c = DS; d = DD;
- a + b + c + d = M – número máximo de pares de pontos no conjunto de dados

os seguintes índices para medir o grau de similaridade entre C e P podem ser definidos (quanto mais altos mais similares são os grupos e as partições):

- *Rand Statistic*: $R = (a + d)/M$
- *Jaccard Coefficient*: $J = a/(a + b + c)$
- *Folkes and Mallows index*: $\sqrt{\frac{a}{a+b} * \frac{a}{a+c}}$
- ...

Clustering - Avaliação

Dunn and Dunn-like indices

Tenta identificar clusters compactos e bem separados. O índice é definido pela equação abaixo, para um número específico de grupos:

$$D_{nc} = min_{i=1,\ldots,nc}\{min_{j=i+1,\ldots,nc}\left(\frac{d(c_i, c_j)}{max_{k=1,..,nc}diam(c_k)}\right)\}$$

onde nc é o número de grupos, $d(c_i, c_j)$ é uma função de dissimilaridade entre dois grupos c_i e c_j definida como

$$d(c_i, c_j) = min_{x \in c_i, y \in c_j} d(x, y) \tag{7}$$

e $diam(C)$ é o diâmetro de um grupo, podendo ser definido como:

$$diam(C) = max_{x,y \in C} d(x, y) \tag{8}$$

Altos valores para este índice indicam a presença de cluster compactos e bem separados.

Técnicas de Agrupamento

- Sarajane M. Peres - sarajane@usp.br
- Clodoaldo A. M. Lima - c.lima@usp.br

Escola de Artes, Ciências e Humanidades - EACH
Universidade de São Paulo - USP