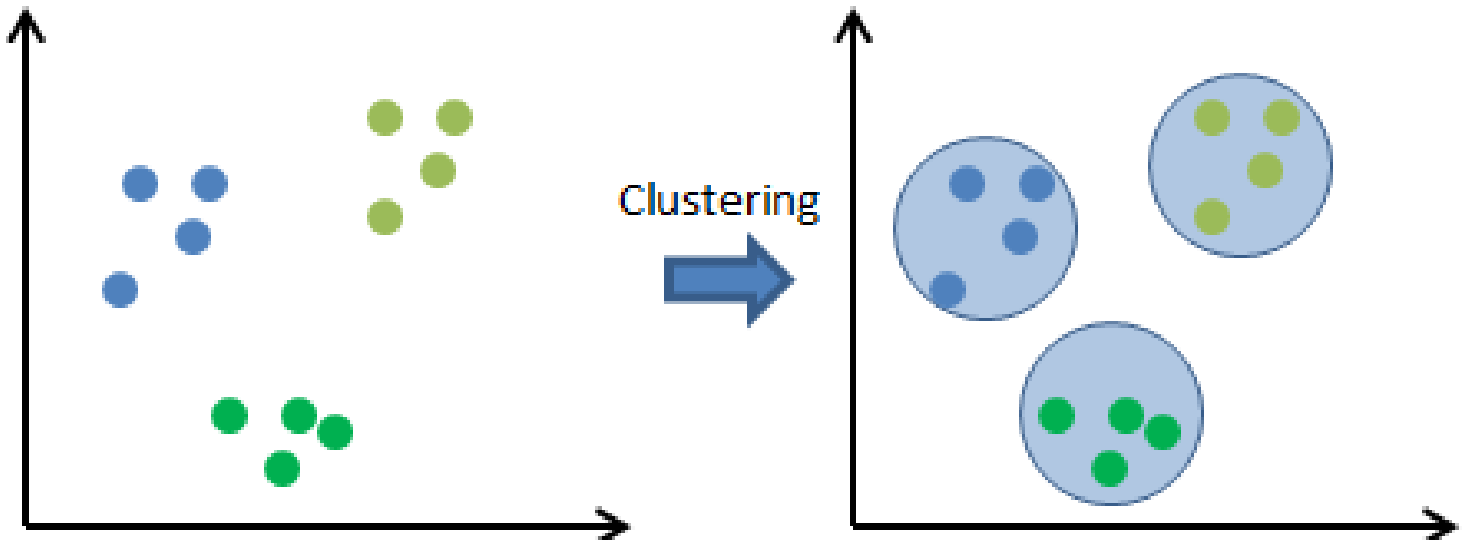
Métodos de Agrupamento para Séries Temporais

Paulo Henrique Sales Guimarães
paulo.guimaraes@ufla.br

Departamento de Estatística/ICET - UFLA

Introdução



Introdução

Vamos criar uma particação da nossa amostra (ou população)

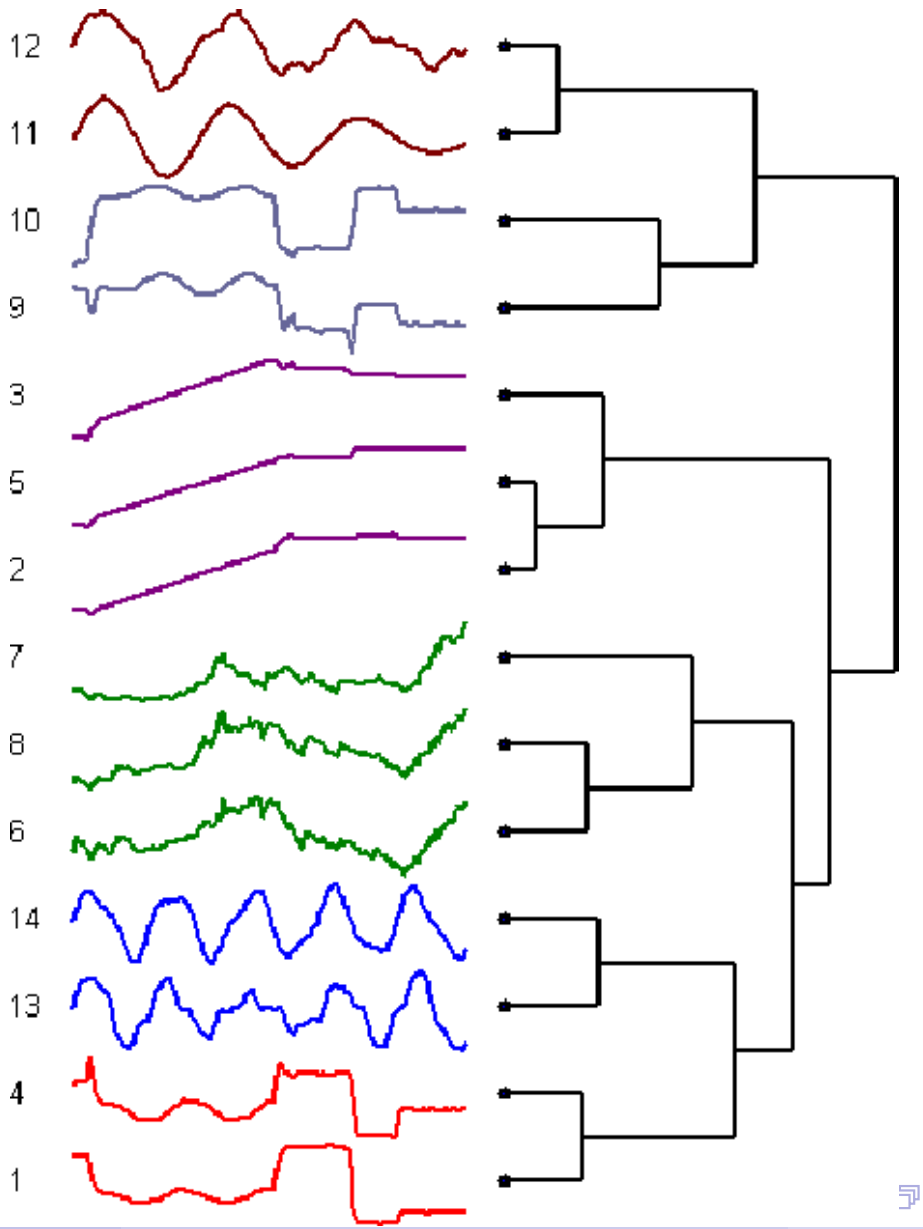
$$C_1, C_2, ..., C_k$$

tal que

$$C_1 \bigcup C_2 \bigcup ... \bigcup C_k = \{1, 2, ..., n\}$$

$$C_i \bigcap C_j = \varnothing, \forall i \neq j.$$

Introdução



Introdução

Exemplos de aplicação

i) Segmentação de mercado (perfil dos consumidores);

Introdução

Exemplos de aplicação

i) Segmentação de mercado (perfil dos consumidores);

ii) Agrupamento baseado em risco de crédito;

Introdução

Exemplos de aplicação

i) Segmentação de mercado (perfil dos consumidores);

ii) Agrupamento baseado em risco de crédito;

iii) Agrupamento por regiões (indicadores macroeconômicos);

Introdução

Exemplos de aplicação

i) Segmentação de mercado (perfil dos consumidores);

ii) Agrupamento baseado em risco de crédito;

iii) Agrupamento por regiões (indicadores macroeconômicos);

iv) Clima (compreender o clima requer encontrar padrões na atmosfera e oceano);

Introdução

Exemplos de aplicação

i) Segmentação de mercado (perfil dos consumidores);

ii) Agrupamento baseado em risco de crédito;

iii) Agrupamento por regiões (indicadores macroeconômicos);

iv) Clima (compreender o clima requer encontrar padrões na atmosfera e oceano);

v) A análise de agrupamentos pode também ser usada para padrões na distribuição espacial ou temporal de uma doença;

Introdução

Exemplos de aplicação

i) Segmentação de mercado (perfil dos consumidores);

ii) Agrupamento baseado em risco de crédito;

iii) Agrupamento por regiões (indicadores macroeconômicos);

iv) Clima (compreender o clima requer encontrar padrões na atmosfera e oceano);

v) A análise de agrupamentos pode também ser usada para padrões na distribuição espacial ou temporal de uma doença;

vii) Agrupamento em séries temporais;

Introdução

Exemplos de aplicação

i) Segmentação de mercado (perfil dos consumidores);

ii) Agrupamento baseado em risco de crédito;

iii) Agrupamento por regiões (indicadores macroeconômicos);

iv) Clima (compreender o clima requer encontrar padrões na atmosfera e oceano);

v) A análise de agrupamentos pode também ser usada para padrões na distribuição espacial ou temporal de uma doença;

vii) Agrupamento em séries temporais;

viii) Identificação de *outliers*.

ix) etc.

Introdução

i) Aprendização não Supervisionado;

Introdução

i) Aprendização não Supervisionado;

ii) Técnica exploratória;

Introdução

i) Aprendização não Supervisionado;

ii) Técnica exploratória;

iii) Existência de relações que emergem dos dados;

Introdução

i) Aprendização não Supervisionado;

ii) Técnica exploratória;

iii) Existência de relações que emergem dos dados;

iv) Similaridade (dissimilaridade);

Introdução

i) Aprendização não Supervisionado;

ii) Técnica exploratória;

iii) Existência de relações que emergem dos dados;

iv) Similaridade (dissimilaridade);

v) Diferentes definições de homogeneidade podem conduzir em agrupamentos bastante diferentes.

Critérios de agrupamento

* Compactação - clusters esféricos e/ou bem separados;

Critérios de agrupamento

* Compactação - clusters esféricos e/ou bem separados;

** Encadeamento ou ligação - não é robusto para casos com pouca separação espacial;

Critérios de agrupamento

* Compactação - clusters esféricos e/ou bem separados;

** Encadeamento ou ligação - não é robusto para casos com pouca separação espacial;

*** Separação espacial - distâncias entre clusters.

Etapas na análise de agrupamento

i) Definir os objetivos do estudo;

Etapas na análise de agrupamento

i) Definir os objetivos do estudo;

ii) Identificar variáveis;

Etapas na análise de agrupamento

i) Definir os objetivos do estudo;

ii) Identificar variáveis;

iii) Selecionar indivíduos a serem agrupados;

Etapas na análise de agrupamento

i) Definir os objetivos do estudo;

ii) Identificar variáveis;

iii) Selecionar indivíduos a serem agrupados;

iv) Preparação dos dados (normalizações, redução de dimensionalidade, *missing values*, entre outros);

Etapas na análise de agrupamento

i) Definir os objetivos do estudo;

ii) Identificar variáveis;

iii) Selecionar indivíduos a serem agrupados;

iv) Preparação dos dados (normalizações, redução de dimensionalidade, *missing values*, entre outros);

v) Proximidade (métricas de distância);

Etapas na análise de agrupamento

i) Definir os objetivos do estudo;

ii) Identificar variáveis;

iii) Selecionar indivíduos a serem agrupados;

iv) Preparação dos dados (normalizações, redução de dimensionalidade, *missing values*, entre outros);

v) Proximidade (métricas de distância);

vi) Agrupamento (por exemplo, partições, hierarquias de partições e partições fuzzy);

Etapas na análise de agrupamento

i) Definir os objetivos do estudo;

ii) Identificar variáveis;

iii) Selecionar indivíduos a serem agrupados;

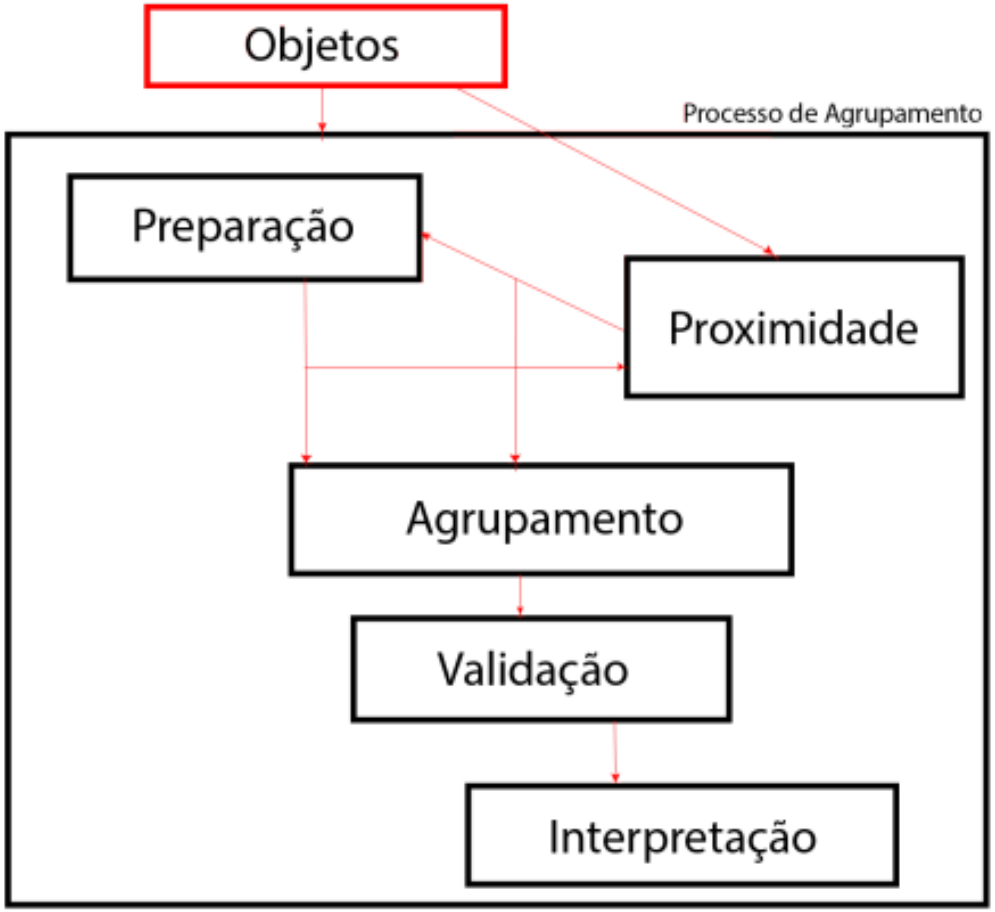
iv) Preparação dos dados (normalizações, redução de dimensionalidade, *missing values*, entre outros);

v) Proximidade (métricas de distância);

vi) Agrupamento (por exemplo, partições, hierarquias de partições e partições fuzzy);

vii) Validação e interpretação dos resultados.

Etapas na análise de agrupamento



Métricas de dissimilaridade

Um dos passos mais importantes durante o processo de agrupamento é a aplicação das medidas de similaridade/distância.

Aplicar as medidas de similaridade sobre séries temporais pode produzir resultados insatisfatórios devido à dependência temporal entre as observações.

Métricas de dissimilaridade

Para determinada métrica ser considerada uma distância ela deve possuir, necessariamente, as seguintes propriedades para vetores $\mathbf{x}, \mathbf{y}$ e $\mathbf{z}$:

Métricas de dissimilaridade

Para determinada métrica ser considerada uma distância ela deve possuir, necessariamente, as seguintes propriedades para vetores $\mathbf{x}, \mathbf{y}$ e $\mathbf{z}$:

* não negatividade: $dist(\mathbf{x}, \mathbf{y}) > 0, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$;
* reflexividade: $dist(\mathbf{x}, \mathbf{y}) = 0$, se, e somente se, $\mathbf{x} = \mathbf{y}$;
* simetria: $dist(\mathbf{x}, \mathbf{y}) = dist(\mathbf{y}, \mathbf{x})$;
* desigualdade triangular: $dist(\mathbf{x}, \mathbf{z}) + dist(\mathbf{z}, \mathbf{y}) \geq dist(\mathbf{x}, \mathbf{y})$.

Métricas de dissimilaridade

A distância de Minkowski é definida como

$$dist(\mathbf{x}, \mathbf{y}) = \left(\sum_{i}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}.$$

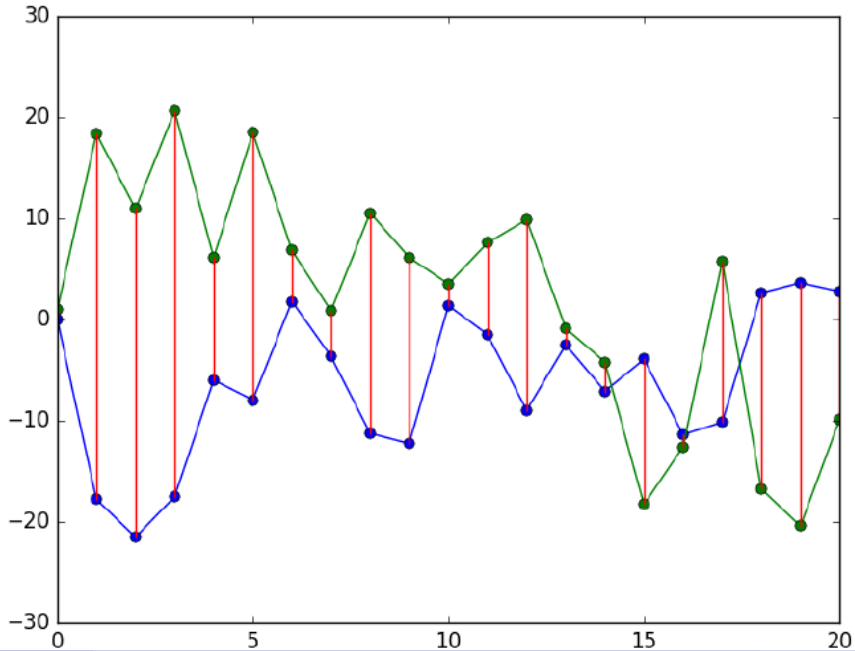
Para $p = 2$ temos a distância Euclidiana ou L_2 e para $p = 1$ a distância Manhattan ou L_1.

Métricas de dissimilaridade

A distância de Minkowski é definida como

$$dist(\mathbf{x}, \mathbf{y}) = \left(\sum_{i}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}.$$

Para $p = 2$ temos a distância Euclidiana ou L_2 e para $p = 1$ a distância Manhattan ou L_1.



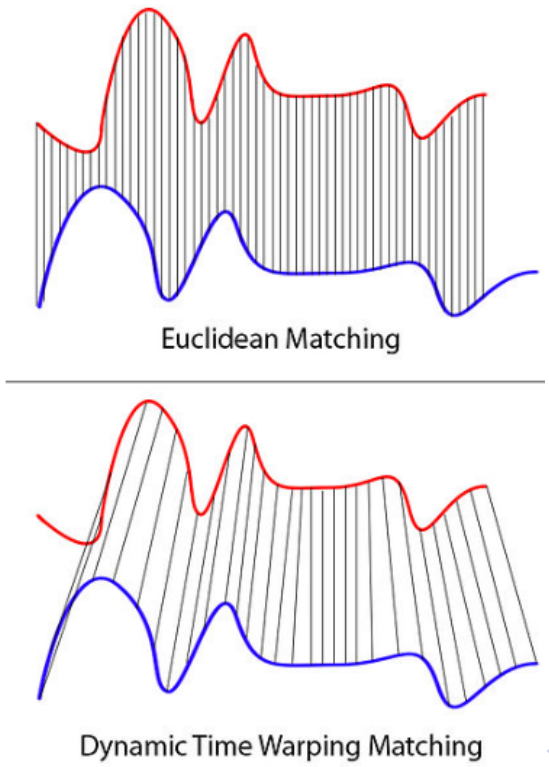
Métricas de dissimilaridade - DTW

A DTW (*Dynamic Time Warping*) é a métrica de dissimilaridade mais utilizada para de séries temporais.

Métricas de dissimilaridade - DTW

A DTW (*Dynamic Time Warping*) é a métrica de dissimilaridade mais utilizada para de séries temporais.

Quanto mais próximo o DTW estiver de zero, maior será a similaridade entre as séries temporais.



Métricas de dissimilaridade - DTW

DTW é um algoritmo que compara e alinha duas séries temporais encontrando o alinhamento não linear ótimo entre esses séries.

Métricas de dissimilaridade - DTW

DTW é um algoritmo que compara e alinha duas séries temporais encontrando o alinhamento não linear ótimo entre esses séries.

Esse algoritmo cria uma matriz de distâncias $m \times m$, sendo m o tamanho das séries, no qual cada ponto na matriz representa a distância euclidina entre dois pontos pontos das séries x e y.

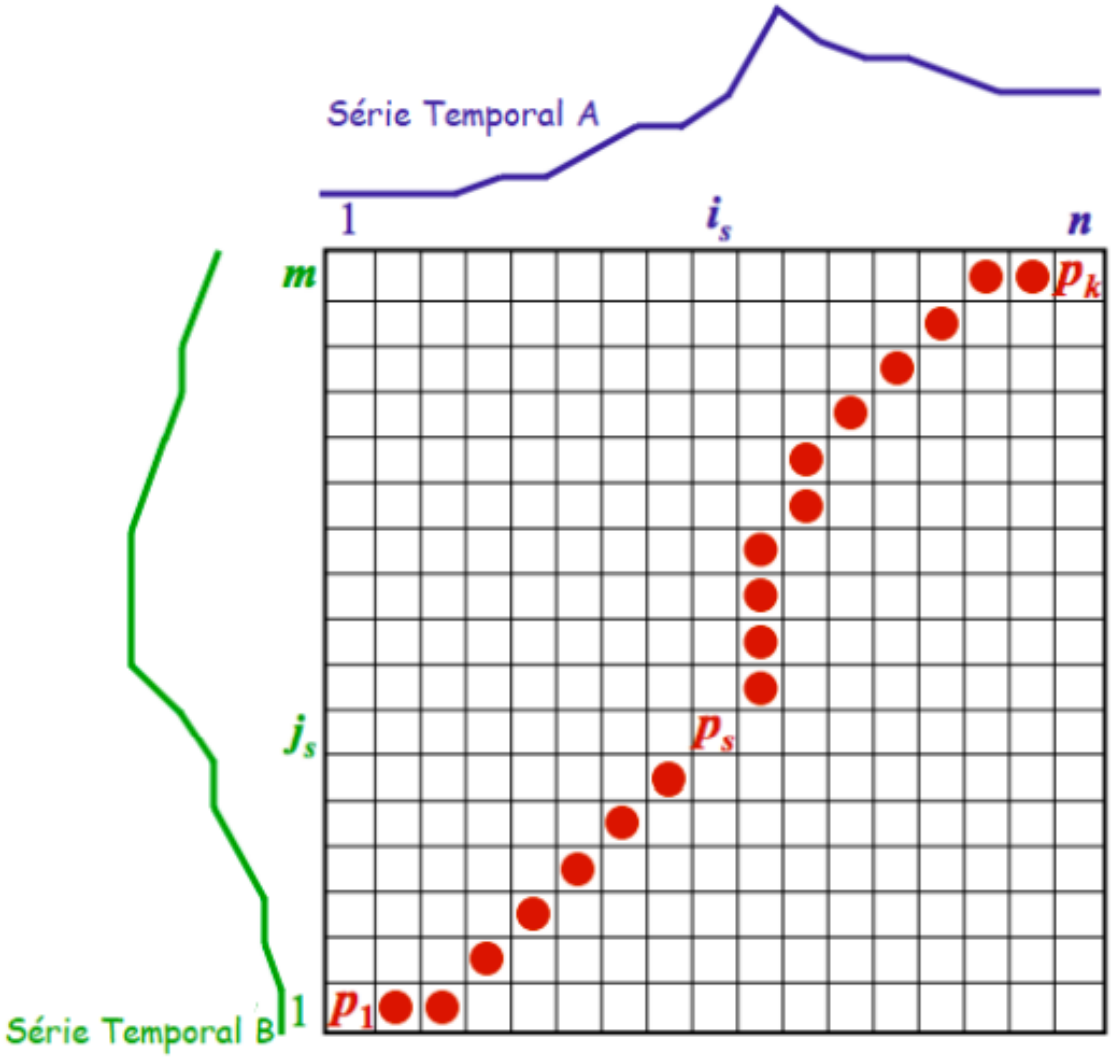
Métricas de dissimilaridade - DTW

DTW é um algoritmo que compara e alinha duas séries temporais encontrando o alinhamento não linear ótimo entre esses séries.

Esse algoritmo cria uma matriz de distâncias $m \times m$, sendo m o tamanho das séries, no qual cada ponto na matriz representa a distância euclidina entre dois pontos pontos das séries x e y.

Um conjunto contínuo $W = \{w_1, w_2, ..., w_k\}$ de elementos da matriz, chamado de *warping path*, com $k \geq m$ define o mapiamento entre x e y.

Métricas de dissimilaridade - DTW



Métricas de dissimilaridade - DTW

Assim, cada w_k corresponde à alguma posição $(i,j)_k$ do grid, tal que a distância DTW entre as séries **x** e **y**, definida por esse caminho seja minimizada:

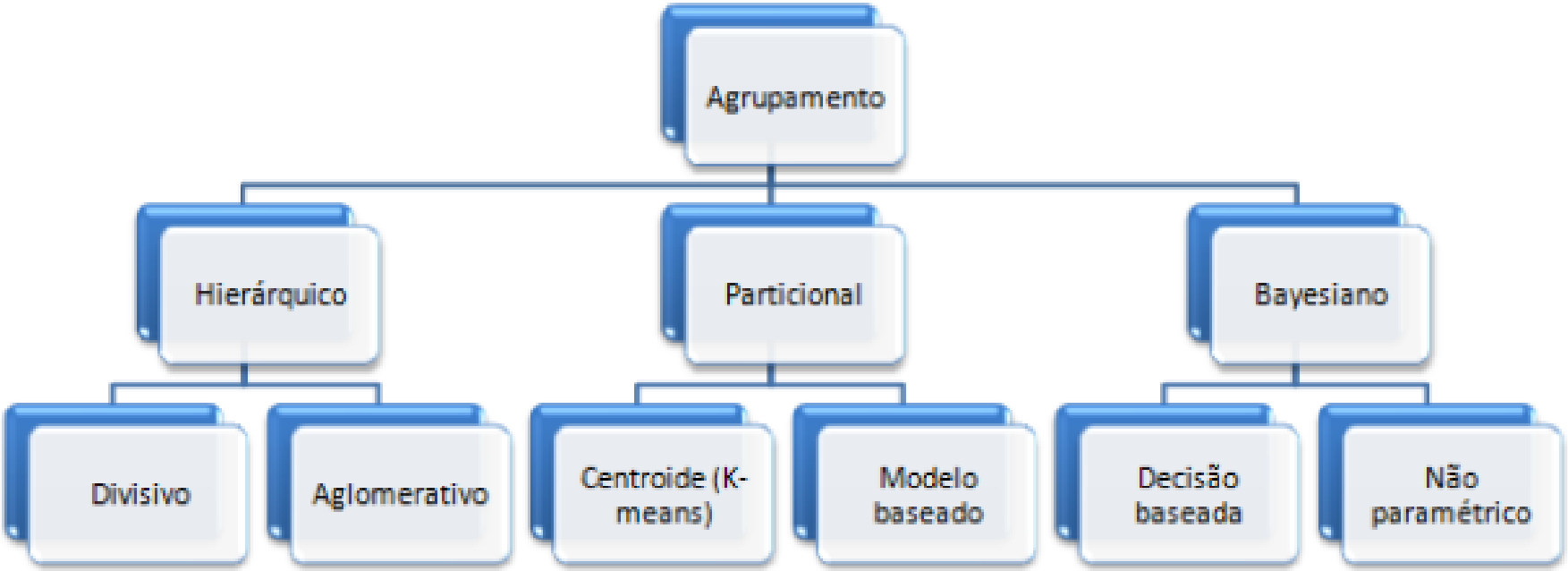
$$DTW(\mathbf{x},\mathbf{y}) = min_W[\sum_{k=1}^{p}\delta(w_k)]$$

.

Desta forma, o cálculo da dissimilaridade DTW pode ser definido por meio da programação dinâmica da variável $\gamma(i,j)$ que representa a soma acumulada da distância escolhida na posição (i,j) do grid:

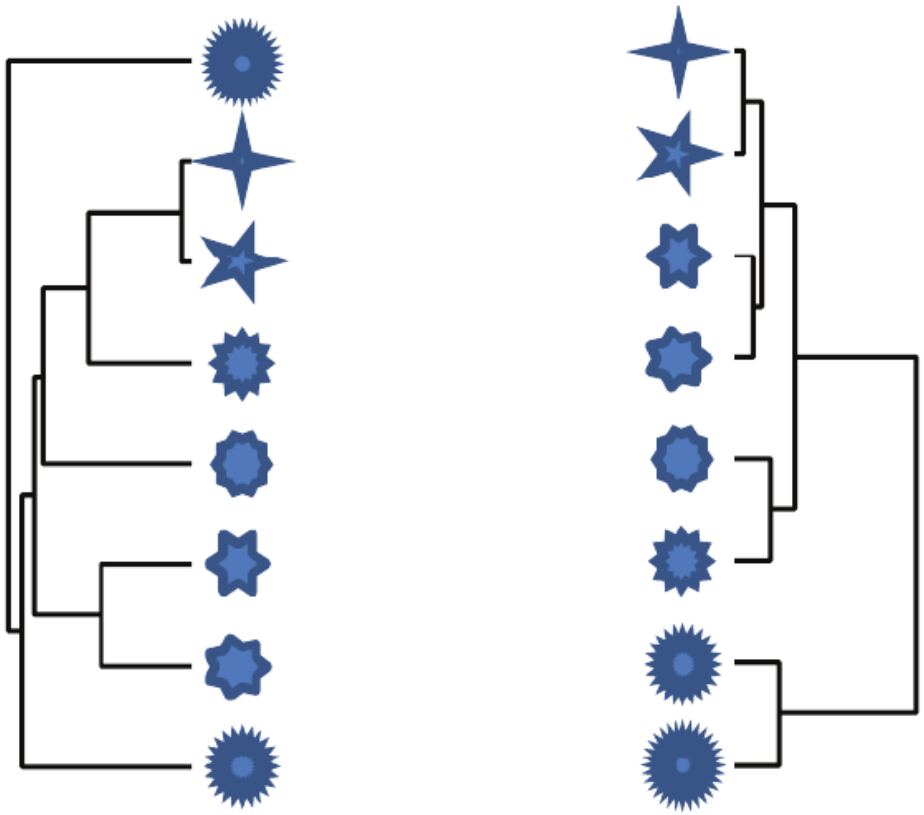
$$\gamma(i,j) = \delta(i,j) + min[\gamma(i-1,j), \gamma(i-1,j-1), \gamma(i,j-1)].$$

Principais algoritmos de agrupamento



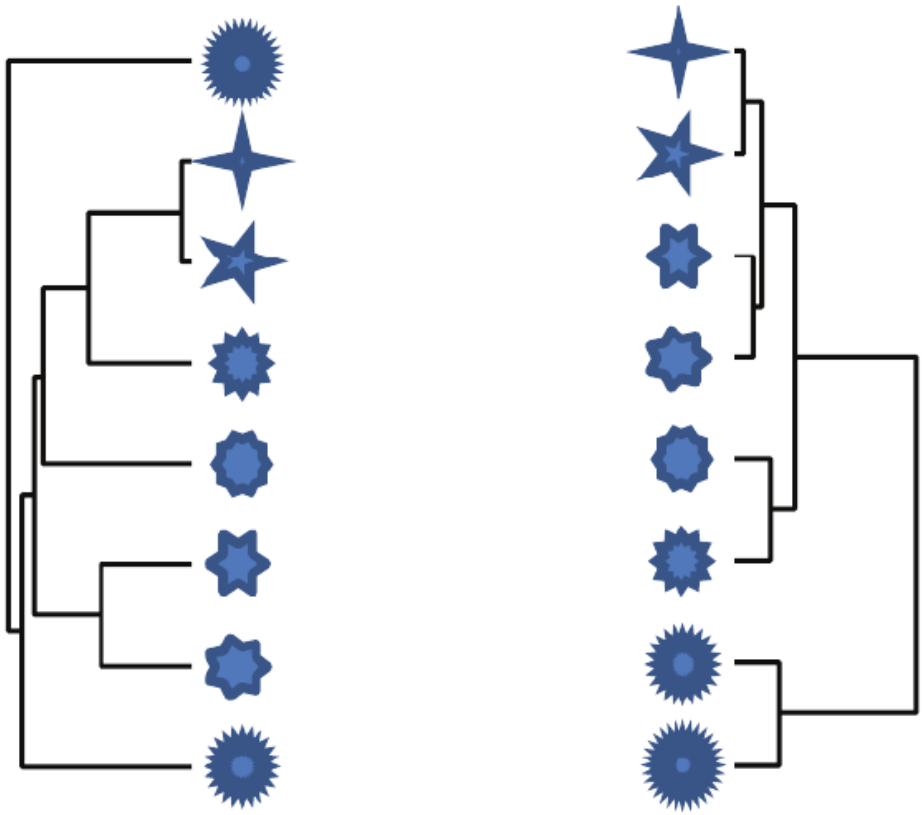
Principais algoritmos de agrupamento

i) Algoritmos hierárquicos ou aglomerativos,;



Principais algoritmos de agrupamento

i) Algoritmos hierárquicos ou aglomerativos,;



ii) Algoritmos particionais ou de atribuição de pontos

Algoritmos hierárquicos

i) Não é necessário informar a prior o número de grupos que os dados se encontram divididos;

Algoritmos hierárquicos

i) Não é necessário informar a prior o número de grupos que os dados se encontram divididos;

ii) Ordem de complexidade $O(n^2 \log(n))$;

Algoritmos hierárquicos

i) Não é necessário informar a prior o número de grupos que os dados se encontram divididos;

ii) Ordem de complexidade $O(n^2 \log(n))$;

iii) Matrizes de características ou de dissimililaridade;

Algoritmos hierárquicos

i) Não é necessário informar a prior o número de grupos que os dados se encontram divididos;

ii) Ordem de complexidade $O(n^2 \log(n))$;

iii) Matrizes de características ou de dissimililaridade;

iv) Divisivos ou aglomerativos;

Algoritmos hierárquicos

i) Não é necessário informar a prior o número de grupos que os dados se encontram divididos;

ii) Ordem de complexidade $O(n^2 \log(n))$;

iii) Matrizes de características ou de dissimililaridade;

iv) Divisivos ou aglomerativos;

v) Dendograma.

Algoritmos hierárquicos

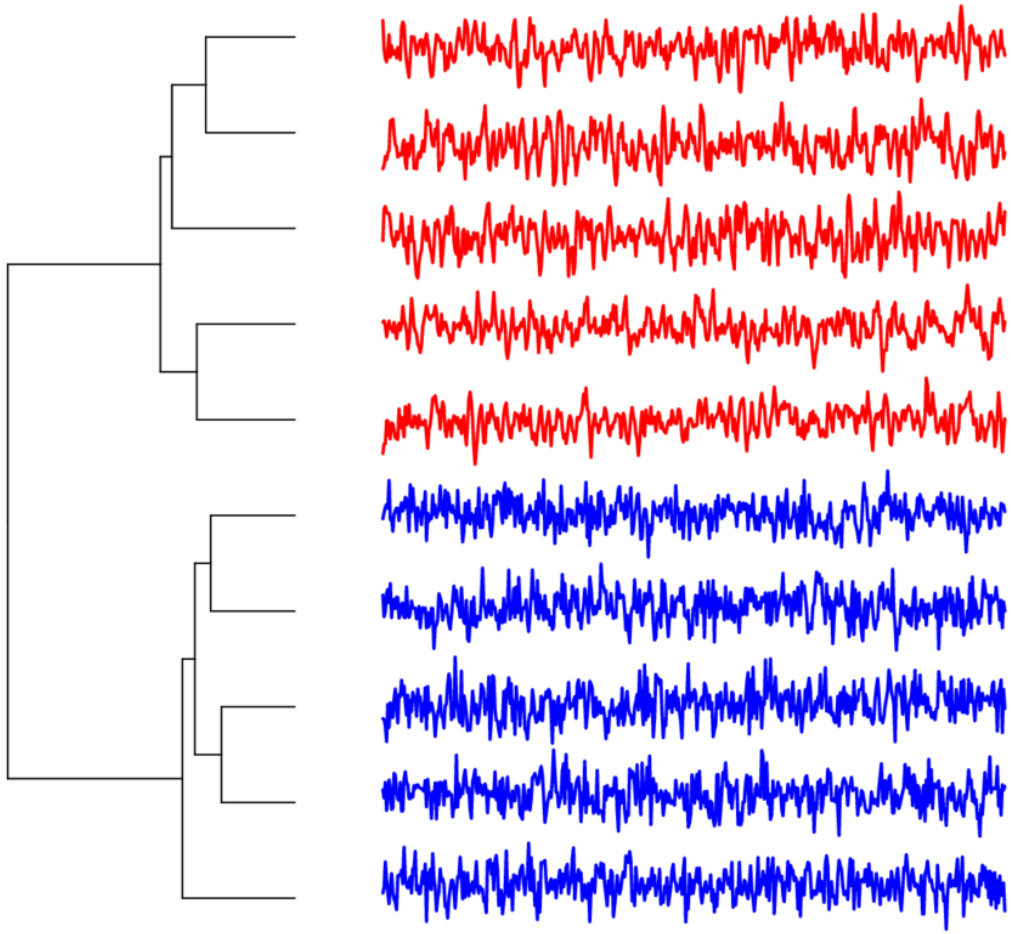


Figure: Exemplo de Dendograma.

Agrupamento de partição K-médias

i) Agrega os pontos em k grupos - soma dos quadrados das k distâncias dos pontos ao centros dos clusters é minimizada;

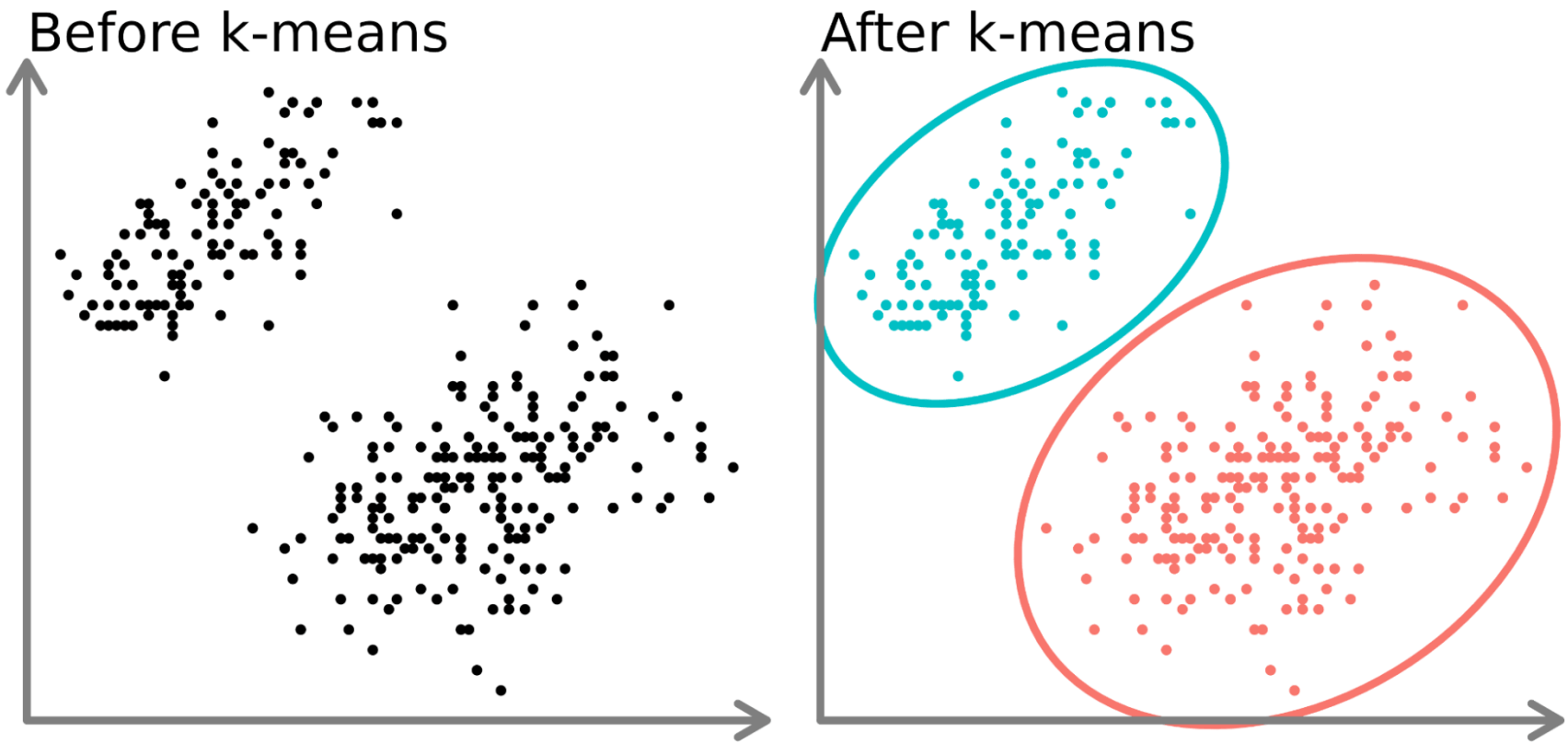
ii) Método baseado em centroides;

iii) Requer que o espaço seja euclidiano;

iv) Algoritmo mais comum foi proposto por Hartigan e Wong em 1979;

v) $W(G_k) = \sum_{x_i \in G_k} \|x_i - \mu_i\|^2$, em que x_i é um ponto no grupo G_k e μ_i é o centróide correspondente.

Agrupamento de partição K-médias



Agrupamento de partição K-médias

Inicialização:

i) Determina-se o número K de *clusters*;

Agrupamento de partição K-médias

Inicialização:

i) Determina-se o número K de *clusters*;

ii) Definimos uma partição inicial (os indivíduos a serem agrupados são alocados em K *clusters*);

Agrupamento de partição K-médias

Inicialização:

i) Determina-se o número K de *clusters*;

ii) Definimos uma partição inicial (os indivíduos a serem agrupados são alocados em K *clusters*);

iii) Determinamos os centróides desses K *clusters*.

Agrupamento de partição K-médias

Inicialização:

i) Determina-se o número K de *clusters*;

ii) Definimos uma partição inicial (os indivíduos a serem agrupados são alocados em K *clusters*);

iii) Determinamos os centróides desses K *clusters*.

Iteração:

i) Realocação: cada indivíduo é alocado ao *cluster* de cujo centroide ele é mais próximo;

Agrupamento de partição K-médias

Inicialização:

i) Determina-se o número K de *clusters*;

ii) Definimos uma partição inicial (os indivíduos a serem agrupados são alocados em K *clusters*);

iii) Determinamos os centróides desses K *clusters*.

Iteração:

i) Realocação: cada indivíduo é alocado ao *cluster* de cujo centroide ele é mais próximo;

ii) Atualização: após a realocação dos indivíduos, os centroides dos novos *clusters* são calculados;

Agrupamento de partição K-médias

Inicialização:

i) Determina-se o número K de *clusters*;

ii) Definimos uma partição inicial (os indivíduos a serem agrupados são alocados em K *clusters*);

iii) Determinamos os centróides desses K *clusters*.

Iteração:

i) Realocação: cada indivíduo é alocado ao *cluster* de cujo centroide ele é mais próximo;

ii) Atualização: após a realocação dos indivíduos, os centroides dos novos *clusters* são calculados;

iii) Retornamos ao passo i.

Agrupamento de partição K-médias

Inicialização:

i) Determina-se o número K de *clusters*;

ii) Definimos uma partição inicial (os indivíduos a serem agrupados são alocados em K *clusters*);

iii) Determinamos os centróides desses K *clusters*.

Iteração:

i) Realocação: cada indivíduo é alocado ao *cluster* de cujo centroide ele é mais próximo;

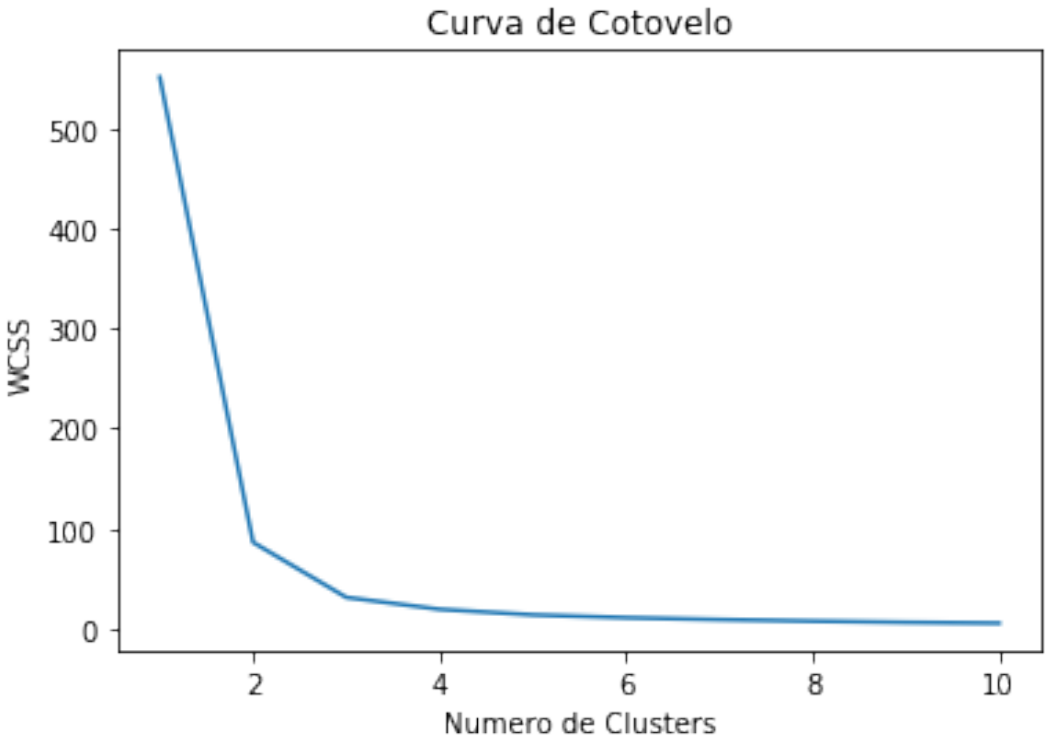
ii) Atualização: após a realocação dos indivíduos, os centroides dos novos *clusters* são calculados;

iii) Retornamos ao passo i.

O processo termina quando nenhum indivíduo for realocado.

Agrupamento de partição K-médias

A curva de cotovelo ou Método *Elbow Curve* é uma técnica usada para encontrar a quantidade ideal de *clusters* K. Este método testa a variância dos dados em relação ao número de clusters. O valor ideal de K é aquele que tem um menor *Within Sum of Squares* (WSS) e ao mesmo tempo o menor número de *clusters*.



Agrupamento de partição K-médias

Algumas limitações do algoritmo K-médias:

i) O número de *clusters* deve ser fixado a priori;

Agrupamento de partição K-médias

Algumas limitações do algoritmo K-médias:

i) O número de *clusters* deve ser fixado a priori;

ii) Só devemos utilizar variáveis quantitativas;

Agrupamento de partição K-médias

Algumas limitações do algoritmo K-médias:

i) O número de *clusters* deve ser fixado a priori;

ii) Só devemos utilizar variáveis quantitativas;

iii) Sensível à presença de *outliers*;

Agrupamento de partição K-médias

Algumas limitações do algoritmo K-médias:

i) O número de *clusters* deve ser fixado a priori;

ii) Só devemos utilizar variáveis quantitativas;

iii) Sensível à presença de *outliers*;

iv) Não é adequado quando a estrutura natural dos *clusters* tem formas não convexas ("não esféricas").

Agrupamento de partição K-medoides

É uma variação do K-médias, na qual a única diferença é que não se faz o uso de um ponto médio, ou centróide, das instâncias de um mesmo grupo, mas sim de um ponto mediano, ou medóide, que melhor representa as instâncias de cada grupo.

Agrupamento de partição K-medoides

É uma variação do K-médias, na qual a única diferença é que não se faz o uso de um ponto médio, ou centróide, das instâncias de um mesmo grupo, mas sim de um ponto mediano, ou medóide, que melhor representa as instâncias de cada grupo.

Diferentemente do K-médias, o K-medoides consegue trabalhar com qualquer métrica de dissimilaridade, além da distância euclidiana.

Agrupamento de partição K-Medoides (Particionamento em Medoides (PAM))

É uma variação do K-médias, na qual a única diferença é que não se faz o uso de um ponto médio, ou centróide, das instâncias de um mesmo grupo, mas sim de um ponto mediano, ou medoide, que melhor representa as instâncias de cada grupo.

Agrupamento de partição K-Medoides (Particionamento em Medoides (PAM))

É uma variação do K-médias, na qual a única diferença é que não se faz o uso de um ponto médio, ou centróide, das instâncias de um mesmo grupo, mas sim de um ponto mediano, ou medoide, que melhor representa as instâncias de cada grupo.

Diferentemente do K-médias, o K-medoides consegue trabalhar com qualquer métrica de dissimilaridade, além da distância euclidiana.

Agrupamento de partição K-Medoides (Particionamento em Medoides (PAM))

É uma variação do K-médias, na qual a única diferença é que não se faz o uso de um ponto médio, ou centróide, das instâncias de um mesmo grupo, mas sim de um ponto mediano, ou medoide, que melhor representa as instâncias de cada grupo.

Diferentemente do K-médias, o K-medoides consegue trabalhar com qualquer métrica de dissimilaridade, além da distância euclidiana.

Da mesma forma que fizemos com o algoritmo K-médias, precisamos também definir um valor de K, isto é, o número de *clusters* e uma partição inicial.

Agrupamento difuso (fuzzy)

A lógica fuzzy ou lógica nebulosa (ou ainda, difusa), surge por meio de sua formulação da teoria dos conjuntos para gerar um certo "afrouxamento"da rigidez numérica da matemática clássica. Esse "afrouxamento" traz praticidade para modelagens matemáticas, por exemplo.

Agrupamento difuso (fuzzy)

A lógica fuzzy ou lógica nebulosa (ou ainda, difusa), surge por meio de sua formulação da teoria dos conjuntos para gerar um certo "afrouxamento"da rigidez numérica da matemática clássica. Esse "afrouxamento" traz praticidade para modelagens matemáticas, por exemplo.

A lógica difusa tem por objetivo modelar modos de raciocínio aproximados ao invés de precisos.

Agrupamento difuso (fuzzy)

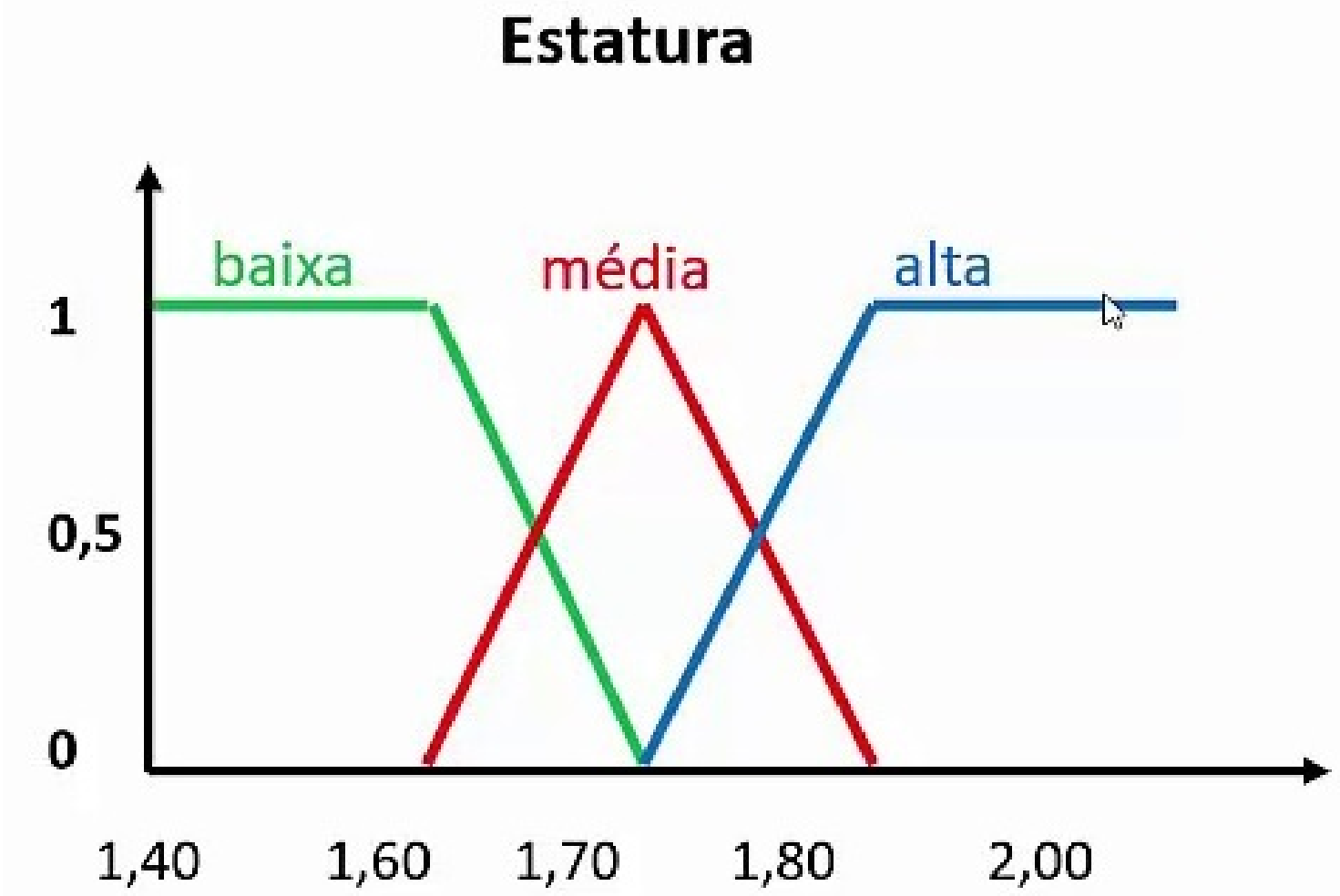
A lógica fuzzy ou lógica nebulosa (ou ainda, difusa), surge por meio de sua formulação da teoria dos conjuntos para gerar um certo "afrouxamento"da rigidez numérica da matemática clássica. Esse "afrouxamento" traz praticidade para modelagens matemáticas, por exemplo.

A lógica difusa tem por objetivo modelar modos de raciocínio aproximados ao invés de precisos.

Na lógica difusa as proposições podem ter valores intermediários entre "Verdadeiro" e "Falso". A veracidade destas é uma função que pode assumir qualquer valor entre 0 (absolutamente falso) e 1 (absolutamente verdadeiro).

Agrupamento difuso (fuzzy)

Agrupamento difuso (fuzzy)



Agrupamento difuso (fuzzy)

O agrupamento difuso (*fuzzy clustering*) é uma forma de agrupamento em que cada elemento pode pertencer a mais de um grupo (*cluster*).

Agrupamento difuso (fuzzy)

O agrupamento difuso (*fuzzy clustering*) é uma forma de agrupamento em que cada elemento pode pertencer a mais de um grupo (*cluster*).

No agrupamento não difuso (também conhecido como agrupamento rígido, ou *hard clustering*), os dados são divididos em grupos distintos, no qual cada ponto de dados só pode pertencer a exatamente um grupo.

Agrupamento difuso (fuzzy)

O agrupamento difuso (*fuzzy clustering*) é uma forma de agrupamento em que cada elemento pode pertencer a mais de um grupo (*cluster*).

No agrupamento não difuso (também conhecido como agrupamento rígido, ou *hard clustering*), os dados são divididos em grupos distintos, no qual cada ponto de dados só pode pertencer a exatamente um grupo.

No agrupamento fuzzy, os pontos de dados podem pertencer a vários clusters. Por exemplo, uma maçã pode ser vermelha ou verde (agrupamento rígido), mas uma maçã também pode ser vermelha E verde (agrupamento suave). Neste exemplo, a maçã pode ser vermelha até certo ponto, assim como verde até certo ponto.

Agrupamento difuso (fuzzy)

Um dos algoritmos de agrupamento difuso mais utilizados é o *Fuzzy C-means clustering* (FCM).

Agrupamento difuso (fuzzy)

Um dos algoritmos de agrupamento difuso mais utilizados é o *Fuzzy C-means clustering* (FCM).

i) O algoritmo de agrupamento difuso **C-means** é muito parecido com o algoritmo de agrupamento K-médias;

Agrupamento difuso (fuzzy)

Um dos algoritmos de agrupamento difuso mais utilizados é o *Fuzzy C-means clustering* (FCM).

i) O algoritmo de agrupamento difuso **C-means** é muito parecido com o algoritmo de agrupamento K-médias;

ii) Escolher o número de *clusters*;

Agrupamento difuso (fuzzy)

Um dos algoritmos de agrupamento difuso mais utilizados é o *Fuzzy C-means clustering* (FCM).

i) O algoritmo de agrupamento difuso **C-means** é muito parecido com o algoritmo de agrupamento K-médias;

ii) Escolher o número de *clusters*;

iii) Atribuir coeficientes aleatoriamente para cada ponto de dados para estar no cluster;

Agrupamento difuso (fuzzy)

Um dos algoritmos de agrupamento difuso mais utilizados é o *Fuzzy C-means clustering* (FCM).

i) O algoritmo de agrupamento difuso **C-means** é muito parecido com o algoritmo de agrupamento K-médias;

ii) Escolher o número de *clusters*;

iii) Atribuir coeficientes aleatoriamente para cada ponto de dados para estar no cluster;

iv) Calcular o centroide para cada *cluster*;

Agrupamento difuso (fuzzy)

Um dos algoritmos de agrupamento difuso mais utilizados é o *Fuzzy C-means clustering* (FCM).

i) O algoritmo de agrupamento difuso **C-means** é muito parecido com o algoritmo de agrupamento K-médias;

ii) Escolher o número de *clusters*;

iii) Atribuir coeficientes aleatoriamente para cada ponto de dados para estar no cluster;

iv) Calcular o centroide para cada *cluster*;

v) Para cada ponto de dados calcular o coeficiente para que esteja nos *clusters*.

Índices de validação de agrupamentos

As diferentes técnicas de agrupamento disponíveis são heurísticas e fornecem uma aproximação para o resultado ideal, podendo gerar resultados diferentes a partir de um único conjunto de dados de entrada.

Índices de validação de agrupamentos

As diferentes técnicas de agrupamento disponíveis são heurísticas e fornecem uma aproximação para o resultado ideal, podendo gerar resultados diferentes a partir de um único conjunto de dados de entrada.

Verificar quantitativamente a diferença entre a aproximação e a resposta real é uma tarefa complexa pois depende do conhecimento do domínio, da aplicação e das técnicas de agrupamento utilizadas.

Índices de validação de agrupamentos

As diferentes técnicas de agrupamento disponíveis são heurísticas e fornecem uma aproximação para o resultado ideal, podendo gerar resultados diferentes a partir de um único conjunto de dados de entrada.

Verificar quantitativamente a diferença entre a aproximação e a resposta real é uma tarefa complexa pois depende do conhecimento do domínio, da aplicação e das técnicas de agrupamento utilizadas.

Um índice de validação indica a qualidade de um determinado agrupamento. Seu cálculo pode ser realizado considerando diferentes funções como o erro quadrático ou a compactação existente entre elementos de um mesmo grupo.

Índices de validação de agrupamentos

Na validação de agrupamento, os índices podem ser classificados conforme três critérios:

i) relativo - busca encontrar um agrupamento que melhor se ajuste aos dados (Dunn [Dunn 1974] e Silhueta [Rousseeuw 1987]);

Índices de validação de agrupamentos

Na validação de agrupamento, os índices podem ser classificados conforme três critérios:

i) relativo - busca encontrar um agrupamento que melhor se ajuste aos dados (Dunn [Dunn 1974] e Silhueta [Rousseeuw 1987]);

ii) interno - avaliam a qualidade de um agrupamento baseando-se apenas nos dados originais (Estatística Gap [Tibshirani, Walther e Hastie 2001]);

Índices de validação de agrupamentos

Na validação de agrupamento, os índices podem ser classificados conforme três critérios:

i) relativo - busca encontrar um agrupamento que melhor se ajuste aos dados (Dunn [Dunn 1974] e Silhueta [Rousseeuw 1987]);

ii) interno - avaliam a qualidade de um agrupamento baseando-se apenas nos dados originais (Estatística Gap [Tibshirani, Walther e Hastie 2001]);

iii) externo - calculam a qualidade de um agrupamento em função de uma estrutura referência (Rand [Rand 1971]; Jaccard [Jaccard 1908]).

Índices de validação de agrupamentos

Vamos considerar que *a* represente o total de pares de objetos que pertencem ao mesmo grupo em α e β;

b representa o total de pares de objetos colocados no mesmo grupo na partição α e em grupos separados na partição β;

c representa o total de pares de objetos colocados em grupos diferentes na partição α e em mesmo grupo na partição β;

d denota o total de pares de objetos colocados em grupos diferentes na partição α e em β.

Índices de validação de agrupamentos

$$Rand(\pi^e, \pi^r) = \frac{(a+d)}{(a+b+c+d)}$$

$$Jaccard(\pi^e, \pi^r) = \frac{a}{(a+b+c)}$$

$$FolkesMallows(\pi^e, \pi^r) = \frac{a}{\sqrt{(a+b)(a+c)}}.$$

Índices de validação de agrupamentos

O índice de Jaccard calcula a probabilidade de dois exemplos pertencerem ao mesmo grupo ou fazerem parte de dois grupos diferentes nas partições.

Índices de validação de agrupamentos

O índice de Jaccard calcula a probabilidade de dois exemplos pertencerem ao mesmo grupo ou fazerem parte de dois grupos diferentes nas partições.

Tanto o índice de Jaccard quanto o índice de Folkes e Mallows calculam a probabilidade de um ou mais exemplos pertencerem ao mesmo grupo e diferentes partições.

Índices de validação de agrupamentos

O índice de Jaccard calcula a probabilidade de dois exemplos pertencerem ao mesmo grupo ou fazerem parte de dois grupos diferentes nas partições.

Tanto o índice de Jaccard quanto o índice de Folkes e Mallows calculam a probabilidade de um ou mais exemplos pertencerem ao mesmo grupo e diferentes partições.

O índice de Jaccard é uma versão estendida e aprimorada do índice de Folkes e Mallows para bases de dados desbalanceadas ou semisupervisionadas.

Índices de validação de agrupamentos

O índice de Jaccard calcula a probabilidade de dois exemplos pertencerem ao mesmo grupo ou fazerem parte de dois grupos diferentes nas partições.

Tanto o índice de Jaccard quanto o índice de Folkes e Mallows calculam a probabilidade de um ou mais exemplos pertencerem ao mesmo grupo e diferentes partições.

O índice de Jaccard é uma versão estendida e aprimorada do índice de Folkes e Mallows para bases de dados desbalanceadas ou semisupervisionadas.

Tanto o índice de Jaccard quanto o índice de Folkes e Mallows trabalham no intervalo $[-1, 1]$, sendo valores mais próximos de 1 melhores para o agrupamento.

Índices de validação de agrupamentos

O silhouette é um índice de validação proposto que tenta avaliar o quão compacta cada instância se encontra dentro de seu próprio grupo, e, simultaneamente, o quão distante, ela é da instância mais próxima, ou mais similar, à ela e que não pertence ao seu mesmo grupo.

Índices de validação de agrupamentos

O silhouette é um índice de validação proposto que tenta avaliar o quão compacta cada instância se encontra dentro de seu próprio grupo, e, simultaneamente, o quão distante, ela é da instância mais próxima, ou mais similar, à ela e que não pertence ao seu mesmo grupo.

Sendo a agora, denotado pela média da métrica de dissimilaridade entre a instância x_i e as demais instâncias contidas em C_k e b o valor mínimo dentre todas as dissimilaridades entre x_i e as demais instâncias não contidas em C_k. O valor de silhouette S_i do ponto x_i é definido por:

$$S_i = \frac{(b-a)}{max(a,b)},$$

em que $a = \frac{1}{n}\sum_j^{|C_k|} dist(x_i, x_j); x_i, x_j \in C_k$ e $b = min(dist(x_i, x_j)); x_i \in C_k, x_j \notin C_k$, sendo $dist(x_i, x_j$ o valor da métrica de dissimilaridade entre as instâncias x_i e x_j.

Índices de validação de agrupamentos

Os valores de S_i variam de 1 a 1, sendo o valor de 1 geralmente quando a atribuição da instância x_i ao seu grupo é equivocada, e 1 quando é acertada. Valores próximos de 0 podem indicar sobreposição de grupos.

Valores próximos de 1 indicam uma boa partição, ao passo que valores próximos de 1 indicam uma partição ruim.

Índices de validação de agrupamentos

Os valores de S_i variam de 1 a 1, sendo o valor de 1 geralmente quando a atribuição da instância x_i ao seu grupo é equivocada, e 1 quando é acertada. Valores próximos de 0 podem indicar sobreposição de grupos.

Valores próximos de 1 indicam uma boa partição, ao passo que valores próximos de 1 indicam uma partição ruim.

O índice de Dunn é definido como a razão entre a mínima distância intra-grupo e a máxima intergrupo. Varia de $[0, \infty]$ e valores menores implicam em partições melhores.

Referências

a) Assunção, J.V.C. Uma Breve Introdução à Mineração de Dados: Bases Para a Ciência de Dados, com Exemplos em R. Novatec Editora, São Paulo, 2021.

b) Faceli, K., Lorena, A.C., Gama, J., Almeida, T. A., Carvalho, A. C.P.L.F. Inteligência Artifical. Uma abordagem de Aprendizado de Máquina. 2ª ed. Rio de Janeiro, LTC, 2021.

c) Gonzaga, S.T. Curso de Séries Temporais. 2019. Disponível em: http://sillasgonzaga.com/material/curso$_s$*eries*$_t$*emporais*/*clusterizacao*

d) Maharaj, E.A., D'Urso, P., Caiado, J. Time Series Clustering and Classification. CRC Press. 2019.

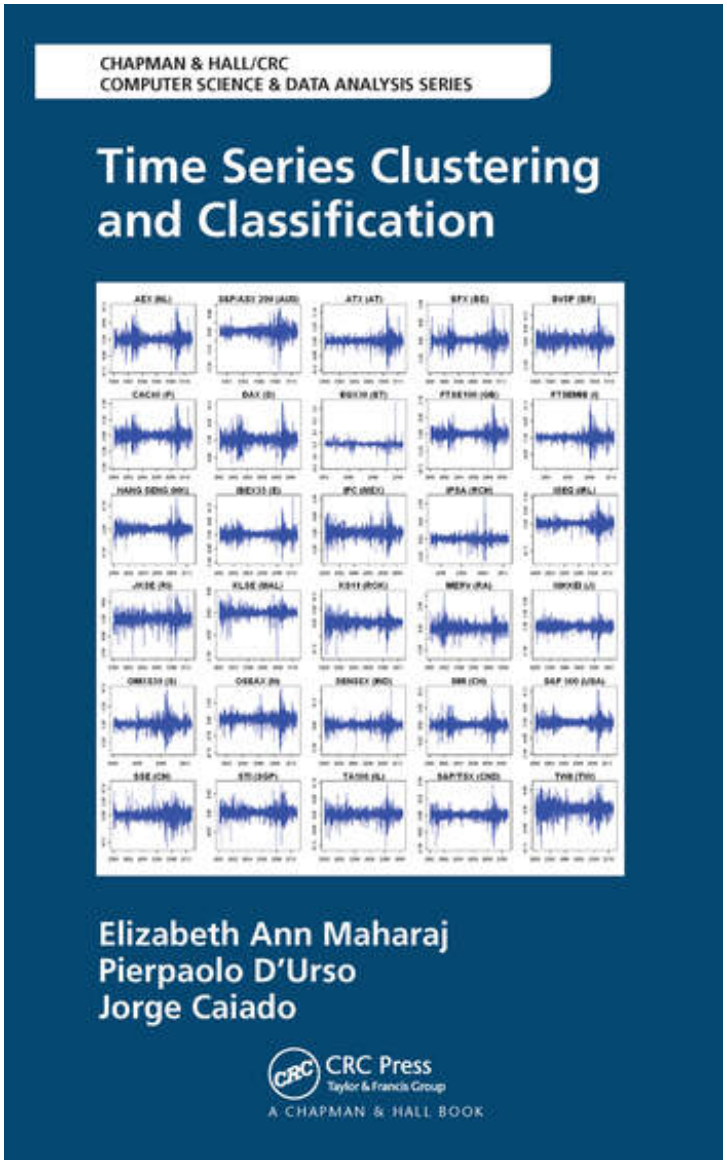
Referências

e) Negri, R.C. Reconhecimento de Padrões. Um estudo dirigido. Blucher, São Paulo, 2021.

e) Sicsú, A.L., Samartini, A., Barth, N.L. Técnicas de Machine Learning. Blucher, São Paulo, 2023.

f) Silva, Mirlei Moura da. Agrupamento de Séries Temporais Utilizando Decomposição de Componentes Estocásticos e Determinísticos. Dissertação. Universidade Federal da Bahia, 2018. 74p.

Referências



Referências

Journal of Statistical Software

November 2014, Volume 62, Issue 1. *http://www.jstatsoft.org/*

TSclust: An R Package for Time Series Clustering

Pablo Montero
University of A Coruña

José A. Vilar
University of A Coruña

Abstract

Time series clustering is an active research area with applications in a wide range of fields. One key component in cluster analysis is determining a proper dissimilarity measure between two data objects, and many criteria have been proposed in the literature to assess dissimilarity between two time series. The R package **TSclust** is aimed to implement a large set of well-established peer-reviewed time series dissimilarity measures, including measures based on raw data, extracted features, underlying parametric models, complexity levels, and forecast behaviors. Computation of these measures allows the user to perform clustering by using conventional clustering algorithms. **TSclust** also includes a clustering procedure based on p values from checking the equality of generating models, and some utilities to evaluate cluster solutions. The implemented dissimilarity functions are accessible individually for an easier extension and possible use out of the clustering context. The main features of **TSclust** are described and examples of its use are presented.

Keywords: time series data, clustering, dissimilarity measure, validation indices.

16º MGEST

Referências

Journal of Statistical Software

August 2009, Volume 31, Issue 7. *http://www.jstatsoft.org/*

Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package

Toni Giorgino
University of Pavia

Abstract

Dynamic time warping is a popular technique for comparing time series, providing both a distance measure that is insensitive to local compression and stretches and the warping which optimally deforms one of the two input series onto the other. A variety of algorithms and constraints have been discussed in the literature. The **dtw** package provides an unification of them; it allows R users to compute time series alignments mixing freely a variety of continuity constraints, restriction windows, endpoints, local distance definitions, and so on. The package also provides functions for visualizing alignments and constraints using several classic diagram types.