AGRUPAMENTO DE SÉRIES TEMPORAIS E MÉTRICAS DE DISSIMILARIDADE - UM ESTUDO EMPÍRICO

Pedro Leal Pazzini da Silva* Cristiano Leite de Castro* Frederico G. Guimarães*

*Programa de Pós-Graduação em Engenharia Elétrica - Universidade Federal de Minas Gerais
Av. Antônio Carlos 6627, 31270-901, Belo Horizonte, MG, Brasil

Email: pedropazzini@ufmg.br crislcastro@ufmg.br fredericoguimaraes@ufmg.br

Abstract— This work presents an empirical analysis on Time Series Clustering. The goal was to check if there is a combination of dissimilarity metric and clustering algorithm that is more relevant than the most commonly used combination: k-means and the Euclidean distance.Three traditional clustering algorithms in combination with four dissimilarity metrics were tested on 55 datasets. Based on significance statistical tests, our results show that the choice of the clustering algotithm can effect the individual performance of a dissimilarity metric. It is also shown that metrics robust to distortions in data, such as DTW and CID, can improve the quality of the resulting partitions.

Keywords— Time series, clustering, dissimilarity metrics, DTW, CID, euclidean distance.

Resumo— Este trabalho apresenta resultados de uma análise empírica no contexto do agrupamento de séries temporais. O objetivo foi verificar se existe alguma combinação, envolvendo algoritmo de agrupamento e métrica de dissimilaridade, que seja mais promissora que a combinação padrão: *k-means* e distância Euclidiana. Foram avaliados 3 algoritmos tradicionais de agrupamento, e suas variações, em combinação com 4 métricas de dissimilaridade sobre 55 bases de dados. Com base em testes estatísticos de significância, os resultados mostraram que a escolha do algoritmo de agrupamento afeta o desempenho individual de uma medida de dissimilaridade e que métricas robustas a distorções nos dados, tais como DTW e CID, podem melhorar significativamente a qualidade das partições obtidas.

Palavras-chave— Séries temporais, agrupamento, métricas de similaridade, DTW, CID, distância euclidiana.

1 INTRODUÇÃO

Séries temporais representam a variação de uma medida de interesse no tempo como, por exemplo, o preço de ações, sinais de eletrocardiogramas, dados meteorológicos, entre outros. Com o advento da Internet das Coisas (do inglês *Internet of Things* ou IoT), é esperado um aumento significativo na disponibilidade de dados na forma de séries temporais, já que sensores de todos os tipos irão transmitir continuamente dados com uma componente temporal associada. Dentre os vários tipos de processamento possíveis a serem realizados com este volume de dados, o *agrupamento de séries temporais* é uma das tarefas mais relevantes, podendo trazer conhecimento útil acerca das relações entre os dados.

Um dos principais desafios em *agrupamento de series temporais* é a escolha apropriada de uma métrica de dissimilaridade. A escolha mais comum é a distância Euclidiana, a qual tem se mostrado competitiva quando o agrupamento é orientado pelo formato das séries (via comparações locais). Em determinados domínios, no entanto, as séries a serem comparadas podem apresentar distorções, tais como diferenças de amplitude e fase, deslocamento local (warping), oclusão, complexidade, entre outras (Batista et al., 2011). Nesse caso, para que seja possível agrupá-las corretamente, é necessário remover as distorções ou lançar mão de uma métrica de dissimilaridade mais robusta, invariante a distorções.

Em um extenso estudo empírico realizado em (Giusti and Batista, 2013), os autores mostraram o desempenho inferior da distância Euclidiana em relação a outras métricas de dissimilaridade, tais como DTW (do inglês *Dynamic Time Warping*) (Berndt and Clifford, 1994) e CID (do inglês *Complexity-Invariant Distance*) (Batista et al., 2011). Este estudo, no entanto, considerou a tarefa de classificação de séries temporais usando como classificador base o algoritmo k-NN (do inglês *k-Nearest Neighbours*). Em se tratando de agrupamento, poucos estudos na literatura têm abordado questões mais gerais relacionadas à comparação de métricas e algoritmos de agrupamento. Estudo recente de (Paparrizos and Gravano, 2015) propõe uma nova medida de dissimilaridade para séries temporais. Na comparação desta com outras tradicionais, os autores argumentaram que o desempenho individual de uma dada medida pode ser influenciado pelo tipo algoritmo de agrupamento adotado.

Do ponto de vista da escolha da métrica e algoritmo para um dado problema de agrupamento orientado à forma, uma pergunta importante a ser formulada é a seguinte: existe alguma combinação, envolvendo algoritmo e medida de dissimilaridade que seja mais apropriada do que a bem conhecida combinação *k-means* e distância Euclidiana? Para tentar responder essa pergunta, nosso estudo empírico avaliou o desempenho de 3 algoritmos tradicionais de agrupamento e, algumas de suas variações, em combinação com 4 métricas de dissimilaridade sobre 55 bases de dados. Usando

a distância Euclidiana como base de comparação, foram selecionadas métricas de dissimilaridade robustas a distorções, tais como DTW e CID, além da métrica CORT que agrega estrutura temporal (correlação) ao processo de comparação das séries. Com relação aos algoritmos de agrupamento, considerou-se na avaliação os bem conhecidos algoritmos *k-means*, *k-medoids* e Agrupamento Hieráquico com diferentes *linkagens*.

Com base em testes estatísticos de significância, os resultados de nossa análise empírica mostraram que a escolha do algoritmo de agrupamento afeta o desempenho individual de uma medida de dissimilaridade (tal como a distância Euclidiana), concordando com o exposto em (Paparrizos and Gravano, 2015). Além disso, nosso estudo aponta que a inclusão do critério de complexidade de séries temporais (CID) pode melhorar significativamente a qualidade das partições obtidas durante o agrupamento.

O restante do artigo encontra-se organizado da seguinte maneira: na Seção 2 serão discutidas as métricas de dissimilaridades usadas nos experimentos. A Seção 3 apresenta o critério adotado para avaliar a qualidade das partições resultantes do agrupamento. Os procedimentos metodológicos e os resultados de nosso estudo empírico são detalhados na Seção 4. Finalmente, as conclusões são expostas na Seção 5.

2 Métricas de dissimilaridade de séries temporais

Métricas de dissimilaridade de séries temporais se enquadram, basicamente, em dois grandes grupos (Montero and Vilar, 2014). O primeiro grupo é orientado ao formato das séries e por isso, leva em conta comparações locais para a descoberta de perfis geométricos. Já o segundo grupo é formado por métricas que consideram estruturas globais das séries temporais, geralmente após uma etapa de pré-processamento ou de extração de características. Medidas estruturais buscam descobrir se as séries possuem dinâmica similar, isto é, se seus crescimentos (positivo ou negativo) a qualquer instanste são semelhantes em direção e taxa. Na análise empírica realizada neste trabalho considerou-se, principalmente, métricas do primeiro grupo, uma vez que as bases de dados selecionadas envolvem problemas de agrupamento orientados ao formato geométrico das séries temporais. Nas seguintes seções as métricas de dissimilaridade consideradas em nosso estudo são brevemente descritas.

2.1 Dynamic Time Warping

Conforme mencionado anteriormente, a distância Euclidiana é a métrica de similaridade mais comumente adotada na literatura tanto para classificação quanto para o agrupamento de séries temporais. Ela é obtida a partir da comparação ponto a ponto entre as séries analisadas. Essa característica, no entanto, pode trazer problemas se as séries apresentarem pequenos deslocamentos em intervalos curtos de tempo (*warpings*).

Dadas duas séries temporais $\boldsymbol{x} = x_1, ..., x_n$ e $\boldsymbol{y} = y_1, ..., y_m$, suponha que seja feito a construção de um *grid* $n_x m$, no qual na posição (i, j) se encontra o valor resultante de uma métrica de dissimilaridade $\delta(i, j)$ entre os pontos x_i e y_j. Tal métrica pode ser, por exemplo, a distância Euclidiana. A partir deste *grid*, é construído um caminho $W = w_1, w_2, w_3, ..., w_k$, onde cada w_k corresponde à alguma posição $(i, j)_k$ do grid, tal que a dissimilaridade DTW entre as séries $\boldsymbol{x}$ e $\boldsymbol{y}$, definida por esse caminho, seja minimizada:

$$DTW(\boldsymbol{x}, \boldsymbol{y}) = min_W[\sum_k \delta(w_k)] \qquad (1)$$

Assim, pode-se definir o cálculo da dissimilaridade DTW por meio da programação dinâmica da variável $\gamma(i, j)$ que representa a soma acumulada da distância escolhida na posição (i, j) do *grid*

$$\gamma(i, j) = \delta(i, j) + min[\gamma(i-1, j), \gamma(i-1, j-1), \gamma(i, j-1)]. \qquad (2)$$

O cálculo se inicia na posição (m, n) e, a partir da fórmula recursiva apresentada na equação 2, constrói-se uma tabela de distâncias acumuladas. O caminho mínimo, que é a dissimilaridade DTW entre as séries temporais, é obtido realizando o caminho reverso, ou seja, com o valor contido em $\gamma(m, n)$.

A ordem de complexidade do algoritmo é $\mathcal{O}(n * m)$. No entanto, a adição de algumas restrições faz com que se alcance ganhos computacionais com pouca perda de performance. Uma alternativa comummente adotada é a criação de uma janela de tamanho ω tal que, $|i_k - j_k| \leq \omega$. Note que para $\omega = 1$, a DTW se torna a distância euclidiana. Em (Giusti and Batista, 2013), sugere-se a utilização de uma janela com tamanho de aproximadamente 10% do tamanho da série temporal, e este será o valor utilizado nos experimentos.

2.2 CORT

A correlação temporal entre as séries $\boldsymbol{x}$ e $\boldsymbol{y}$ de tamanho n pode ser obtida por

$$CORT(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_i^{n-1}(x_i - x_{i+1})(y_i - y_{i+1})}{\sqrt{\sum_i^{n-1}(x_i - x_{i+1})^2}\sqrt{\sum_i^{n-1}(y_i - y_{i+1})^2}}, \qquad (3)$$

onde $-1 \leq CORT(\boldsymbol{x}, \boldsymbol{y}) \leq 1$.

Em alguns domínios, é desejável uma métrica que agregue estrutura temporal (correlação) ao processo de comparação local entre as séries. Tal métrica pode ser obtida combinando-se um componente temporal de correlação com uma métrica

convencional orientada à forma como, por exemplo, a distância Euclidiana ou a DTW. A importância do componente de correlação para o cálculo da dissimilaridade é controlada por uma função de sintonia $f(\cdot)$. Caso a correlação entre as séries temporais a serem agrupadas é relevante para a tarefa de agrupamento, então a função de sintonia deve corrigir as dissimilaridades obtidas pela métrica convencional, amplificando (ou reduzindo) seu valor em função da correlação temporal entre as séries. Assim, uma métrica de dissimilaridade orientada à forma e, que também leva em conta o aspecto temporal entre as séries $\boldsymbol{x}$ e $\boldsymbol{y}$ é dada pela seguinte expressão

$$dist(\boldsymbol{x}, \boldsymbol{y}) = f(CORT(\boldsymbol{x}, \boldsymbol{y})).dist_Q(\boldsymbol{x}, \boldsymbol{y}) \quad (4)$$

onde $dist_Q$ é a distância convencional entre as séries e $f(\cdot)$ é a função de sintonia que deve amplificar a distância convencional quando a correlação temporal se encontra no intervalo $[-1, 0[$, reduzi-la quando a correlação temporal se encontra no intervalo $]0, 1]$, e não alterá-la quando a correlação for igual à 0. A seguinte função de sintonia, proposta em (Chouakria and Nagabhushan, 2007), atende a estes critérios e foi utilizada neste trabalho

$$f(z) = \frac{2}{1 + \exp(\gamma z)}, \quad \gamma \geq 0, \quad (5)$$

onde γ é um parâmetro que controla o quanto o aspecto temporal deve influenciar na métrica de dissimilaridade.

2.3 Complexity-Invariant Distance

De uma maneira simples, pode-se definir a complexidade de uma série temporal como o número de picos e vales que ela possui (Batista et al., 2011). Curvas suaves com poucos picos e vales têm complexidade menor do que séries temporais mais ruidosas que possuem muitos picos e vales. Tal estimativa de complexidade pode ser entendida como uma medida do tamanho da série temporal quando ela for esticada. Assim, dada uma série temporal $\boldsymbol{x}$, sua estimativa de complexidade é dada por

$$CE(\boldsymbol{x}) = \sqrt{\sum_{i}^{n-1}(x_i - x_{i+1})^2}. \quad (6)$$

No cálculo da dissimilaridade entre duas séries temporais, uma medida da complexidade relativa entre elas pode ser usada para corrigir as possíveis distorções associadas às diferenças de complexidade. Tais distorções não seriam capturadas caso somente uma métrica convencional, tal como a distância Euclidiana, fosse aplicada. Tal fator de correção e a dissimilaridade invariante à complexidade, entre as séries temporais $\boldsymbol{x}$ e $\boldsymbol{y}$, são dados por

$$CF(\boldsymbol{x}, \boldsymbol{y}) = \frac{max(CE(\boldsymbol{x}), CE(\boldsymbol{y}))}{min(CE(\boldsymbol{x}), CE(\boldsymbol{y}))}, \quad (7)$$

$$dist_{CID} = dist_Q(\boldsymbol{x}, \boldsymbol{y}).CF(\boldsymbol{x}, \boldsymbol{y}), \quad (8)$$

onde $dist_Q$ é uma métrica de dissimilaridade convencional, tais como a distância Euclidiana ou a DTW.

3 Validação do agrupamento

Na literatura, os índices que medem a qualidade da partição obtida por um algoritmo de agrupamento podem ser divididos em índices internos, os quais utilizam informações somente dos dados e de suas dissimilaridades, e em índices externos, que além das informações já mencionadas utilizam dados rotulados, ou a partição esperada, para a contabilização do índice. Os índices de validação externa não possuem aplicação prática em tarefas de agrupamento reais, já que nelas não se possui *a priori* a rotulação dos dados. No entanto, eles são relevantes quando é necessário validar alguma estratégia ou parâmetros do agrupamento.

Um dos índices externos mais usados para medir a qualidade do agrupamento é o índice de Variação de Informação (VI), proposto em (Meilă, 2007). O motivo é que, diferentemente dos demais índices externos, o VI é calculado no espaço de partições via Teoria da Informação. Caso se disponha da partição real dos dados, o VI reflete a qualidade do particionamento obtido por um dado algoritmo de agrupamento em relação a partição real. Tal propriedade será explorada neste artigo.

Dada uma partição $\mathcal{P}$ composta por n instâncias divididas em k grupos, onde o k-ésimo grupo contém n_k instâncias, é possível definir uma variável aleatória como a probabilidade de se escolher uma instância do k-ésimo grupo e a Entropia da partição da seguinte forma

$$P(k) = \frac{n_k}{n}, \quad (9)$$

$$H(\mathcal{P}) = -\sum_{i}^{k} P(k) \log P(k). \quad (10)$$

Se $P(k)$ e $P'(k')$ são as variáveis aleatórias associadas às partições, de uma mesma base de dados, $\mathcal{P}$ e $\mathcal{P}'$ respectivamente, então é possível definir a probabilidade de uma instância pertencer ao grupo $\mathcal{C}_k$ da partição $\mathcal{P}$ e ao grupo $\mathcal{C}'_{k'}$ da partição $\mathcal{P}'$

$$P(k, k') = \frac{|\mathcal{C}_k \cap \mathcal{C}'_{k'}|}{n}, \quad (11)$$

e a informação mútua entre $\mathcal{P}$ e $\mathcal{P}'$ como

$$I(\mathcal{P},\mathcal{P}') = \sum_{i}^{k}\sum_{i}^{k'} P(k,k')\log\frac{P(k,k')}{P(k)P'(k')}. \quad (12)$$

Finalmente, a variação de informação (VI) entre $\mathcal{P}$ e $\mathcal{P}'$ é definida por

$$VI(\mathcal{P},\mathcal{P}') = H(\mathcal{P}) + H(\mathcal{P}') - 2I(\mathcal{P},\mathcal{P}'). \quad (13)$$

O índice VI é mínimo, e igual a zero, quando as partições $\mathcal{P}'$ e $\mathcal{P}$ são idênticas. Seu valor máximo, por sua vez, ocorre quando as partições são completamente aleatórias, não possuindo qualquer informação mútua. Neste caso seu valor é $H(\mathcal{P}) + H(\mathcal{P}')$.

4 Experimentos e resultados

Nossa análise empírica tem por objetivo verificar se existe alguma combinação, envolvendo algoritmo de agrupamento e métrica de dissimilaridade, que seja mais promissora que a combinação padrão, *k-means* e distância Euclidiana, comumente adotada na literatura. Assume-se que não se tem conhecimento *a priori* do processo de geração dos dados e também das possíveis distorções (amplitude, fase, deslocamento local, etc.) que estes possam apresentar. Essa premissa vai de encontro ao que ocorre na maior parte das aplicações práticas que envolvem o agrupamento de séries temporais.

Foram selecionadas 55 bases de dados reais do repositório *UCR Time series*, disponível em (Chen et al., 2015). Todas as séries neste repositório encontram-se previamente normalizadas segundo a transformação Z (Lin and Li, 2009) para eliminar possíveis diferenças de amplitude e *offset* que possam prejudicar sua comparaçãoAdicionalmente, todas as bases de dados possuem rotulação e assim, o índice VI (Seção 3) foi aplicado para medir a qualidade da partição obtida em contraste com a partição esperada. Foram também computados os tempos de processamento de cada estratégia, definidos como a soma do tempo do agrupamento em si, com o tempo de obtenção das matrizes de dissimilaridade.

Testes estatísticos de significância foram aplicados para verificar se existe diferença entre as estratégias avaliadas (Demšar, 2006). Foram considerados os desempenhos obtidos em relação ao índice VI, que mede a qualidade da partição obtida por cada estratégia e também, em relação ao tempo de processamento. Em cada experimento, os *rankings* médios de cada estratégia foram calculados e o teste de Friedman foi aplicado. Dados M conjunto de dados e L algoritmos com seus rankings médios R_t onde $t = 1, ..., L$, assume-se a hipótese nula H_0 de que os algoritmos são equivalentes. Neste caso, a estatistica

$$F_F = \frac{(M-1)\chi_F^2}{M(L-1)-\chi_F^2} \quad (14)$$

é distribuída de acordo com a distribuição F com $L-1$ e $(L-1)(M-1)$ graus de liberdade, onde

$$\chi_F^2 = \frac{12M}{L(L+1)}\left(\sum_t R_t^2 - \frac{L(L+1)^2}{4}\right). \quad (15)$$

Caso a H_0 seja rejeitada pelo teste de Friedman, os testes *post-hoc* de Nemenyi e Bonferroni-Dunn podem ser aplicados para apontar significância na comparação um-contra-um e um-contra-todos, respectivamente.

4.1 Experimento 1 - Algoritmos de agrupamento

O objetivo deste experimento foi verificar o efeito do algoritmo de agrupamento sobre o desempenho individual de uma determinada métrica de dissimilaridade. Caso o algoritmo não exerça qualquer influência, é esperado que as combinações "algoritmo + métrica" sejam equivalentes em termos das partições obtidas. As combinações testadas são mostradas na Tabela 1. Observe que a métrica distância Euclidiana foi fixada, por se tratar de uma métrica comumente usada na literatura, enquanto os algoritmos e suas variações foram variados.

	Id	Algoritmo	Similaridade
0	KDR_eucl.	k-medoids(random)	eucl.
1	HC_eucl.	hierarchical-complete	eucl.
2	KDP_eucl.	k-medoids(PAM)	eucl.
3	KMP_eucl.	k-means(++)	eucl.
4	KMR_eucl.	k-means(random)	eucl.
5	HA_eucl.	hierarchical-average	eucl.

Tabela 1: Estratégias comparadas no Exp. 1.

Para as 55 bases de dados e 6 combinações (estratégias), o valor crítico da estatística de Friedman é de 2.24. O valor médio dos *rankings* forneceu uma estatística $F_F = 12.72$ rejeitando, portanto, H_0. Dessa forma, pode-se afirmar que a escolha do algoritmo de agrupamento tem efeito significativo no desempenho individual de uma dada medida de similaridade.

Finalmente, o teste *post-hoc* de Bonferroni-Dunn foi aplicado, centrado na combinação padrão, *k-means* e distância Euclidiana. Os resultados deste teste (vide Figura 1) mostram que o algoritmo hierárquico com linkagem *average* possui desempenho significativamente superior a estratégia padrão. Além disso, ficou evidenciada a superioridade dos algoritmos hierárquicos em relação aos algoritmos particionais no contexto de séries temporais.

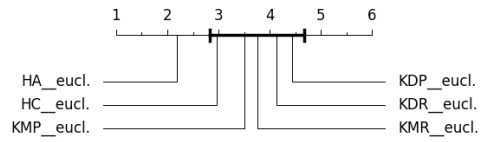

Figura 1: Resultado do teste de Bonferroni após o ranqueamento dos valores de VI.

4.2 Experimento 2 - Métricas de dissimilaridade

O objetivo deste experimento foi comparar o desempenho das diferentes métricas de similaridade, cujas propriedades foram discutidas na Seção 2. Desde que o algoritmo exerce influência no desempenho individual da métrica, o procedimento usado neste experimento foi fixar o algoritmo de agrupamento e variar as métricas. O algoritmo fixado foi o agrupamento hierárquico com linkagem *average*, por ter apresentado o melhor resultado no experimento anterior. A combinação *k-means* e distância Euclidiana foi também avaliada neste experimento, por se tratar da combinação padrão adotada na literatura. Todas as combinações deste experimento são mostradas na Tabela 2.

	Id	Algoritmo	Similaridade
0	HA__cort-eucl.	hierarchical-average	cort-eucl.
1	HA__DTW	hierarchical-average	DTW
2	HA__cid-eucl.	hierarchical-average	cid-eucl.
3	KMR__eucl.	k-means(random)	eucl.
4	HA__eucl.	hierarchical-average	eucl.

Tabela 2: Estratégias comparadas no Exp. 2.

Para 55 bases de dados e 5 combinações testadas, o valor crítico da estatística de Friedman é de 2.41. A hipótese H_0 foi descartada, já que obteve-se $F_F = 14.04$. Em seguida, deu-se prosseguimento ao teste *post-hoc* de Bonferroni-Dunn, centrado na estratégia padrão. O resultado, ilustrado na figura 2, mostra que a combinação, *k-means* e distância Euclidiana, é inferior às demais, e além disso, que o uso da DTW e da CID acarretaram em melhoras significativas em relação à distância Euclidiana e CORT-Euclidiana.

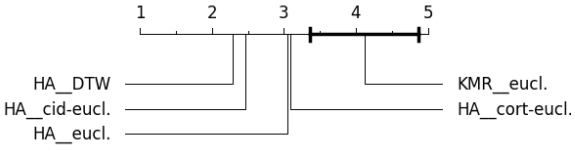

Figura 2: Resultado teste Bonferroni após o ranqueamento dos valores de VI.

4.3 Experimento 3 - Comparação entre Métricas Promissoras

Os experimentos anteriores mostraram que a adição de métricas robustas a distorções nos dados, tais como DTW e CID pode melhorar significativamente a qualidade das partições obtidas. Com isso em mente, o objetivo deste experimento foi comparar as possíveis combinações envolvendo estas métricas promissoras nos quesitos de qualidade das partições (índice VI) e custo computacional (tempo de processamento). As combinações analisadas neste experimento se encontram na Tabela 3. Observe que, novamente, o algoritmo hierárquico com linkagem *average* foi fixado.

	Id	Algoritmo	Similaridade
0	HA__cort-eucl.	hierarchical-average	cort-eucl.
1	HA__DTW	hierarchical-average	DTW
2	HA__cid-eucl.	hierarchical-average	cid-ecul.
3	HA__cort-DTW	hierarchical-average	cort-DTW
4	HA__cid-DTW	hierarchical-average	cid-DTW

Tabela 3: Estratégias do terceiro exeprimento.

O teste de Friedman foi aplicado para testar a hipótese nula de que as estratégias são equivalentes. Para 55 bases de dados e 5 estratégias, o valor crítico é de 2.41. Considerando o quesito tempo de processamento, obteve-se $F_F = 543.03$ e, para o VI, $F_F = 3.48$. Logo, para ambos os casos, a hipótese nula de que os rankings médios são iguais foi descartada. Deu-se então prosseguimento ao teste *post-hoc* de Nemenyi, com o objetivo de verificar quais grupos de estratégias possuem desempenhos equivalentes. O diagrama resultante do teste de Nemenyi para o índice VI pode ser visto na figura 3. O diagrama correlato, para o tempo de processamento, é mostrado na figura 4.

Pode-se concluir, com base na Figura 3, que todas as combinações envolvendo as métricas DTW e CID são significamente melhores que a combinação envolvendo CORT e Euclidiana (HA_cort-eucl.). A Figura 4, por sua vez, mostra que todas as combinações envolvendo DTW (HA-DTW, HA_CID-DTW, HA_cort-DTW) possuem maior custo computacional, sendo estatisticamente inferiores em relação as demais combinações. Uma análise conjunta, envolvendo os dois quesitos, apontam a combinação HA_CID-eucl. como um bom equilíbrio entre a qualidade das partições obtidas e o menor tempo de processamento.

Com relação ao índice VI (Figura 3), observa-se ainda que, embora o teste não tenha apontado diferença estatística, os resultados em termos dos *rankings* médios apontam que a correção CID (complexidade) produz efeito mais positivo do que a correção CORT (correlação temporal), independentemente da métrica convencional adotada: Euclidiana ou DTW. Além disso, todas as combinações envolvendo DTW se mos-

traram melhores que as outras em termos de *rankings* médios, o que sugere a importância de um critério que seja invariante a deslocamentos locais no agrupamento de séries temporais.

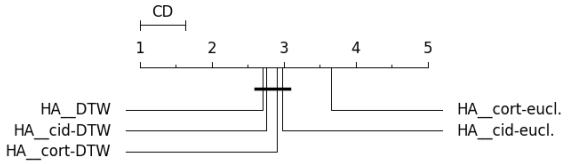

Figura 3: Resultado teste Nemenyi após o ranqueamento dos valores de VI.

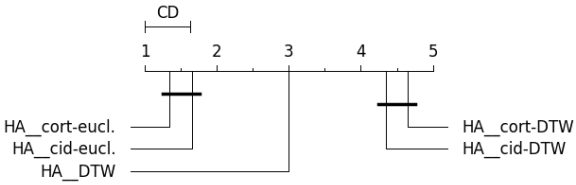

Figura 4: Resultado teste Nemenyi após o ranqueamento dos valores de tempo de processamento.

5 Conclusões

A análise empírica realizada mostrou que de fato, existem estratégias de agrupamento superiores à bem conhecida combinação *k-means* e distância Euclidiana. A adição de métricas robustas a distorções nos dados, tais como DTW e CID, bem como a adição de um componente de correlação temporal (CORT) pode melhorar significativamente a qualidade das partições obtidas. Adicionalemente, nossos resultados mostraram que a escolha do algoritmo de agrupamento tem efeito significativo no desempenho individual das métricas de dissimilaridade, conforme observado em estudo recente de (Paparrizos and Gravano, 2015). Dentre os algoritmos testados neste trabalho, o algoritmo hierárquico com linkagem *average* obteve o melhor desempenho.

No estudo comparativo entre as estratégias mais promissoras, envolvendo combinações entre as métricas CID, DTW, CORT e Euclidiana e, o algoritmo hierárquico, dois quesitos importantes foram analisados: o custo computacional e a qualidade do agrupamento. Se o primeiro quesito não for relevante para a tarefa de agrupamento em questão, nosso estudo recomenda fortemente o uso de DTW, uma vez que todas as estratégias baseadas nesta métrica mostraram-se superiores à combinação: algoritmo hierárquico e Euclidiana. Por outro lado, se o tempo de processamento representar um gargalo, nosso estudo recomenda o uso da combinação algoritmo hierárquico, CID e Euclidiana. Tal combinação mostrou-se estatisticamente equivalente a DTW em termos da qualidade do agrupamento porém, superior em termos de tempo de processamento.

Agradecimentos

O presente trabalho foi realizado com o apoio financeiro da CAPES - Brasil.

Referências

Batista, G. E. A. P. A., Wang, X. and Keogh, E. J. (2011). A complexity-invariant distance measure for time series., *SDM*, SIAM / Omnipress, pp. 699–710.

Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series., *in* U. M. Fayyad and R. Uthurusamy (eds), *KDD Workshop*, AAAI Press, pp. 359–370.

Chen, Y., Keogh, E., Hu, B., Begum, N., Bagnall, A., Mueen, A. and Batista, G. (2015). The ucr time series classification archive. `www.cs.ucr.edu/~eamonn/time_series_data/`.

Chouakria, A. D. and Nagabhushan, P. N. (2007). Adaptive dissimilarity index for measuring time series proximity., *Adv. Data Analysis and Classification* **1**(1): 5–21.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets, *J. Mach. Learn. Res.* **7**: 1–30.

Giusti, R. and Batista, G. E. A. P. A. (2013). An empirical comparison of dissimilarity measures for time series classification, *Proceedings of the 2013 Brazilian Conference on Intelligent Systems*, BRACIS '13, IEEE Computer Society, Washington, DC, USA, pp. 82–88.

Lin, J. and Li, Y. (2009). Finding structural similarity in time series data using bag-of-patterns representation, *Proceedings of the 21st International Conference on Scientific and Statistical Database Management*, SSDBM 2009, Springer-Verlag, Berlin, Heidelberg, pp. 461–477.

Meilă, M. (2007). Comparing clusterings—an information based distance, *J. Multivar. Anal.* **98**(5): 873–895.

Montero, P. and Vilar, J. A. (2014). TSclust: An R package for time series clustering, *Journal of Statistical Software* **62**(1): 1–43.

Paparrizos, J. and Gravano, L. (2015). k-shape: Efficient and accurate clustering of time series, *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, ACM, New York, NY, USA, pp. 1855–1870.