# Cluster-CV: Uma Abordagem de Visão Computacional para a Identificação Espacial de Agrupamentos de Dados

Brayan A. Jaimes, Cristiano L. Castro, Luiz B.Torres*, Gustavo L. Silva, and Antonio P. Braga

Programa de Pós-Graduação em Engenharia Elétrica - Universidade Federal de Minas Gerais - Av. Antônio Carlos 6627, 31270-901, Belo Horizonte, MG, Brasil
{payo.rene,crislcastro,luizlitc,gustavolacerdas}@gmail.com,apbraga@ufmg.br

**Resumo** Este trabalho apresenta uma análise da Matriz de Distâncias através de Visão Computacional (VC) com o objetivo de identificar o número $k$ de agrupamentos sobre bases de dados com sobreposição entre grupos, sendo esta uma abordagem de agrupamento não supervisionado. Assim, com a Matriz de Distâncias que é obtida de uma métrica de distância aplicada par-a-par no conjunto de dados, são aproveitadas e extraídas informações visuais para identificar de forma individual cada um dos agrupamentos contidos nos dados. As amostras pertencentes a cada agrupamento são projetadas em um novo espaço linear, de forma que a sobreposição e distância de separação entre agrupamentos (*clusters*) seja corrigida e aumentada sem perder informação. Os resultados da metodologia aplicada nos experimentos mostram ser promissores, garantindo agrupamentos de dados sem sobreposição e sem perda de informação.

**Keywords:** Matriz de Distâncias, VC, Agrupamento não supervisionado, Sobreposição, Autovetor, Autovalor.

## 1 Introdução

No agrupamento de dados não supervisionado os rótulos das amostras são desconhecidos, não existe informação *a priori* sobre a classe ou grupo onde uma amostra analisada pode ser atribuída. Isto é porque na maioria dos problemas reais a rotulação de dados tem um alto custo computacional e esforço considerável [1]. No entanto, o agrupamento não supervisionado é muito utilizado na mineração de dados, onde o conteúdo de grandes bases de dados não é conhecido [2]. De forma que, neste tipo de abordagem é necessário definir algum critério de dissimilaridade que permita diferenciar ou separar as amostras por grupos [3,4]. Este critério pode ser definido em função de métricas de distância tais como a Euclidiana [5], Minkowski [6], Manhattan [6], Hamming [7] e Mahalanobis [8], sendo as mais utilizadas na literatura.

* Bolsista do CNPq–Brasil (N°150254/2016-4)

As métricas de distância são representadas por uma função $F(x_1, x_2)$ que retorna um valor quantitativo da dissimilaridade entre duas amostras analisadas $x_1$ e $x_2$. O tipo de função de dissimilaridade a utilizar depende do tipo de dados ou a natureza do problema que está sendo analisado. De forma geral, esta função deve apresentar propriedades de não-negatividade, reflexividade, simetria e desigualdade triangular [9].

Além disso, a Matriz de Distâncias é amplamente utilizada na análise de agrupamentos de dados, reconhecimento de padrões, seleção de características, regressão, classificação, entre outras [10]. Dado que esta matriz representa uma das formas mais eficientes de visualização de dissimilaridade nos dados através de uma métrica de distância [10], permitindo assim, obter uma representação visual global das relações de proximidade entre amostras de todo o conjunto de dados. No trabalho de Weiss Y. [11], obtiveram um bom desempenho ao utilizar os autovetores da matriz de dissimilaridade para fazer segmentação em imagens reais e sintéticas. Afirmam que, a partir da inspeção visual na matriz de dissimilaridade, existem informações relevantes da segmentação correta. Por sua vez, em [12] é estudado o agrupamento espectral e como ele determina de forma automática o número de grupos, baseado no cálculo dos autovetores das matrizes de dissimilaridade calculadas entre pares de amostras. No estudo de [13], apresentou-se uma nova abordagem de regularização para ELMs, utilizando informação espacial *a priori* contida na matriz de dissimilaridade. Nesse trabalho foi utilizada a matriz de dissimilaridade de Cossenos que é livre de parâmetros fazendo com que a regularização seja feita sem o ajuste dos mesmos.

Com tudo isto, em um cenário de agrupamento não supervisionado e com existência de sobreposição entre grupos; este trabalho tem como objetivo aproveitar a informação visual fornecida pela Matriz de Distâncias para obter o número $k$ de agrupamentos. Dessa forma, utilizando técnicas de Visão Computacional será possível identificar de forma individual cada um dos agrupamentos contidos nos dados. Depois, através de uma projeção linear feita sobre os agrupamentos de dados, será corrigida a sobreposição e aumentada a distância de separação entre agrupamentos (clusters) sem perder informação.

Este trabalho está distribuído da seguinte forma: na Secção 2 é abordado o problema de agrupamento com sobreposição, onde é evidenciado a necessidade de propor métodos de agrupamento para lidar com a sobreposição de dados. Na Seção 3, é apresentada a metodologia proposta, que utiliza Visão Computacional na análise da Matriz de Distâncias; e o conceito de covariância, autovetor e autovalor para projetar linearmente os agrupamentos sem sobreposição. A discussão e apresentação dos resultados experimentais são mostrados na Seção 4. Finalmente, as conclusões e comentários finais são abordados na Seção 5.

## 2 O Problema de Agrupamento com Sobreposição

Ao longo dos últimos anos, é percebido um aumento considerável no interesse de abordar problemas com sobreposição de dados; pois estes tendem a gerar agrupamentos indesejáveis no conjunto de dados [14]. Uma das explicações deste comportamento é que a maioria dos algoritmos de agrupamento existentes na

literatura são de natureza particional, projetados para agrupar conjuntos de dados bem comportados, rígidos e sem sobreposição [1]. Este é um desafio em métodos de agrupamentos de dados, uma vez que a maioria das bases de dados (reais ou sintéticas) possuem sobreposição [15].

No agrupamento de dados o objetivo principal é encontrar grupos adequados com uma mínima variabilidade intra-grupo (clusters compactos) e máxima separação inter-grupos (clusters bem separados)[16]. Além disso, espera-se que a distância entre agrupamentos seja grande e que o diâmetro esperado dos agrupamentos seja pequeno. A relação dessas duas distâncias é definida como o índice de validação ($val$) definido na Equação (1), e com ele é possível definir o número adequado de agrupamentos [16].

$$val = \frac{Intra}{Inter} = \frac{\frac{1}{N}\sum_{i=1}^{k}\sum_{x\varepsilon C_i} \parallel x - z_i \parallel^2}{min_{i,j}\left(\parallel z_i - z_j \parallel^2\right)} \tag{1}$$

Onde, $i = 1, 2, ..., k-1$; $j = i+1, i+2, ..., k$, $N$ é o número de dados, $k$ é o número de agrupamentos, e $z_i$ representa o centroide do agrupamento $C_i$. No entanto, com a presença de sobreposição os critérios de variabilidade intra-grupo e inter-grupo serão mais difíceis de serem satisfeitos porque a distância de separação inter-grupos é pequena e o diâmetro dos agrupamentos aumenta. Isto ocasiona que o índice de validação ($val$) se torne sensível ao ruído inserido pela sobreposição. Este tipo de situação é evidenciado na Figura 1, onde as regiões de sobreposição são apresentadas na forma de elipses (vermelhas).

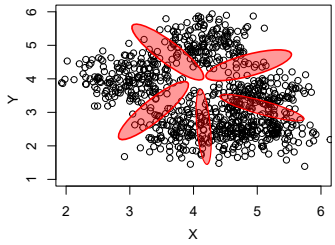



Figura 1: Regiões de sobreposição

## 3 Metodologia proposta

Na Figura 2, é apresentado um fluxograma que explica de forma geral a metodologia proposta por esse trabalho, para o agrupamento de dados não supervisionado baseado na análise da Matriz de Distâncias.

A metodologia proposta consta de duas partes. A primeira etapa consiste na identificação do número de agrupamentos $k$ em que o conjunto de dados pode ser dividido, através da Matriz de Distâncias analisada com Visão Computacional (VC). A segunda parte consiste na projeção linear dos agrupamentos de forma que as distâncias intra-grupos sejam menores e a distância inter-grupos seja maior do que o conjunto de dados original. A seguir, será explicada mais detalhadamente a metodologia e cada uma das etapas.

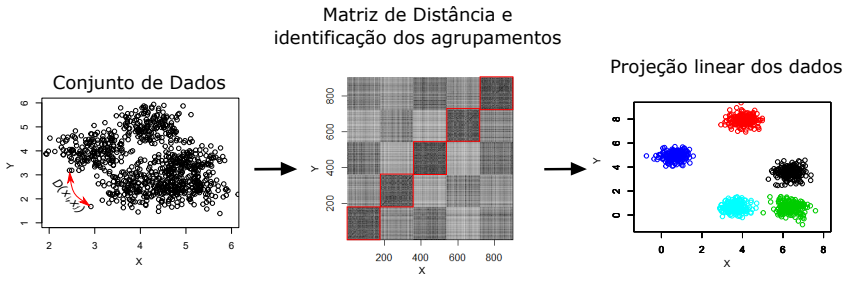



Figura 2: Metodologia proposta

### 3.1 Identificação do Número de Grupos *k*

Uma vez feita a leitura dos dados, é calculada a matriz de distâncias par-a-par no conjunto de dados, utilizando uma métrica de distância que determinará o grau de dissimilaridade entre amostras. Uma das métricas mais intuitiva e utilizada na literatura é a Distância Euclidiana, definida pela Equação (2).

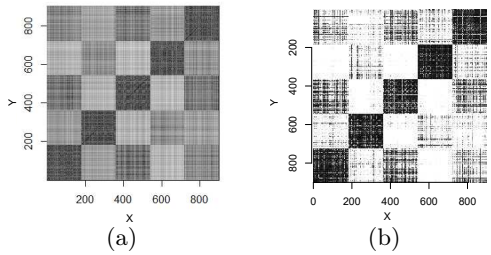
$$D(\mathbf{x_i}, \mathbf{x_j}) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + ... + (x_{in} - x_{jn})^2} \tag{2}$$

A Matriz de Distâncias fornece então uma representação visual geral das relações de proximidade entre todas as amostras, como pode ser observado na Figura 3(a). As regiões mais escuras indicam conjuntos de amostras com uma maior similaridade, ou seja, uma menor distância entre pares de amostras. Nas regiões mais claras indicam que existe uma maior distância entre pares de amostras comparadas. Desta forma, é possível evidenciar que na diagonal secundária existem várias regiões (blocos) que possuem uma maior similaridade que os outros. Estes blocos representam o número de agrupamentos $k$ em que o conjunto de dados pode ser dividido.

A análise de Visão Computacional [17] apresenta uma forma robusta de identificar rapidamente estes agrupamentos. Para isto, a Matriz de Distâncias ($M$) é convertida em imagem, para posterior análise e identificação dos agrupamentos. Inicialmente, a matriz é normalizada na escala de $[0 - 1]$ e é calculado o desvio padrão ($dp$) sobre todos os valores da matriz. Com isto é realizada uma primeira limiarização definida pela Equação (3), sobre a matriz de distâncias com a finalidade de eliminar parcialmente o ruído e regiões que não são de interesse na análise posterior. O limiar utilizado foi de $(1.3 * dp)$, tal valor foi definido com base em experimentos, e representa o 45% da variabilidade do ruído presente na imagem. Tal definição permitiu reduzir o ruído sem degradar as regiões de interesse da imagem. A Figura 3(b) apresenta o resultado após a limiarização.

$$L(x,y) = \begin{cases} 1, & se \quad M(x,y) \geq 1.3 * dp \\ M(x,y), & se \quad M(x,y) < 1.3 * dp \end{cases} \tag{3}$$

A matriz de Distâncias limiarizada $L$ é salva como imagem no formato (PNG - *Portable Network Graphics*). Sobre a imagem é aplicado um filtro gaussiano (Equação (4)) com o objetivo de fazer borramento, reduzir a textura e homogeneizar o ruído presente na imagem. Posteriormente, a imagem é binarizada para

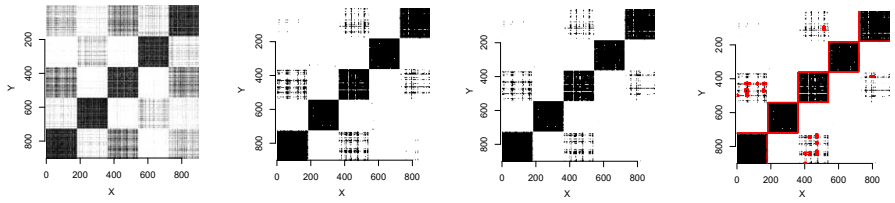



(a) (b)

Figura 3: (a) Matriz de Distâncias; (b) Limiarização da Matriz de Distâncias.

facilitar a busca de agrupamentos [18] e aplicam-se operadores morfológicos de erosão e dilatação sobre a imagem. Por último, são encontrados os contornos dos agrupamentos com operadores morfológicos de preenchimento de regiões. Na Figura 4 é apresentada toda a evolução do processamento realizado na imagem até chegar na identificação dos grupos, através de VC e operadores morfológicos.

$$filtro\ Gaussiano(x,y) = \frac{1}{2\pi\sigma^2} exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \tag{4}$$




(a) (b) (c) (d)

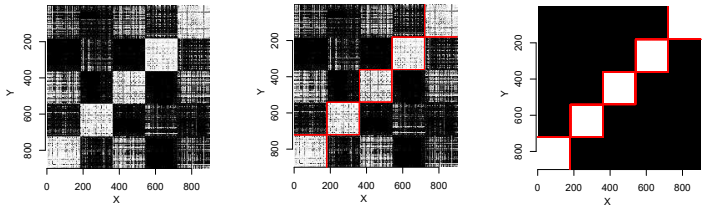
Figura 4: (a) Imagem após do filtro gaussiano; (b) Imagem binarizada; (c) Operadores Morfológicos; (d) Identificação dos agrupamentos.

Outra abordagem para identificar os agrupamentos sobre a imagem analisada é fazer leitura do complemento da imagem $(1 - imagem)$, e sobre ela aplicar os operadores morfológicos de preenchimento de regiões para identificar as regiões de interesse. Na Figura 5 mostram-se os resultados obtidos após a aplicação desta abordagem.

Com os agrupamentos já identificados sobre a imagem, é possível saber especificamente quais amostras dentro do conjunto de dados pertencem a cada agrupamento. Isto é possível porque a relação do número de amostras é equivalente ao número de *pixels* na imagem salva. De forma que, cada amostra no conjunto de dados representa um *pixel* na imagem da Matriz de Distâncias (Figura 5(c)).

### 3.2 Projeção Linear de Agrupamentos

Esta etapa tem como objetivo projetar os agrupamentos de forma que não exista sobreposição entre grupos nem perda de informação, diminuindo a distância intra-grupos e ao mesmo tempo aumentando a distância inter-grupos. Para isto, é necessário identificar o grau de dispersão de cada um dos agrupamentos dentro do




(a) (b) (c)

Figura 5: (a) Complemento da imagem ($1-imagem$); (b) Identificação dos agrupamentos (c) Número de amostras equivalente ao número de *pixels* em cada agrupamento.

conjunto de dados, e baseado nessa informação corrigir a sobreposição existente entre grupos.

Uma forma mais intuitiva de saber o grau de dispersão de um conjunto de dados é com o conceito de variância dos dados. O desvio padrão que é a raiz quadrada da variância proporciona uma medida de quanto os dados estão espalhados no espaço de características $2D$ (Equação (5)). Portanto, a matriz de covariância resume todas as informações das variâncias calculadas em torno dos eixos $x$ e $y$ respectivamente, e a correlação que existe entre as variâncias dos eixos vertical e horizontal (Equação (6)).

$$\sigma_x^2 = \frac{1}{N-1}\left[\sum_{i=1}^{N}(x_i-\mu)^2\right] \qquad (5) \qquad\qquad \Sigma = \begin{bmatrix}\sigma(x,x) & \sigma(x,y)\\ \sigma(y,x) & \sigma(y,y)\end{bmatrix} \qquad (6)$$

Desta forma, o processo de projeção dos agrupamentos no conjunto de dados é definido pelo Algoritmo 1.

**Algorithm 1:** Projeção Linear de Agrupamentos

Dado o número de agrupamento identificados na Matriz de Distâncias através de VC:

**(i)** Para todo o conjunto de dados calcule o centroide $C_T$, e para cada agrupamento identifique o centroide $C_i$.

**(ii)** Calcular a matriz de covariância $\Sigma$ para cada grupo analisado.

**(iii)** Fazer a decomposição da matriz de covariância em auto-vetores e auto-valores.

**(iV)** Projetar a transformação linear $T$ dos agrupamentos, fazendo a redução da distância intra-grupos através de um fator de escala para cada eixo $x$ e $y$, utilizando a matriz de escala $S$. Se for necessário Rotacionar o agrupamento, definido pela matriz $R$:

$$S = \begin{bmatrix}s_x & 0\\ 0 & s_y\end{bmatrix} \qquad R = \begin{bmatrix}\cos(\theta) & -\sin(\theta)\\ \sin(\theta) & \cos(\theta)\end{bmatrix} \qquad (7)$$

De modo que;

$$T = RS \qquad (8)$$

**(V)** Aumentar a separação inter-grupos levando em consideração a posição de cada um dos centroides $C_i$. Garantir que a distância entre os centroides $C_T$ e $C_i$ seja maior que a atual. O termo $q$ é o fator multiplicativo de separação variado pelo usurário, que determina quanto os grupos vão se afastar.

$$d(C_i) = (\sqrt{(C_T - C_j)^2}) * q, \quad q > 1 \qquad (9)$$

## 4 Resultados Experimentais e Análise

Com o objetivo de testar a metodologia proposta e observar o comportamento dos dados resultantes foram realizados três experimentos diferentes. As métricas de avaliação na qualidade do agrupamento e na partição dos grupos obtidos em cada experimento, foram baseadas em índices de qualidade de agrupamento interno como *Silhouette*; e agrupamento externo como *Precision*, *Recall*, *Jaccard* [19,20]. Estes índices de qualidade variam de $[0-1]$, onde resultados próximos de 1 indicam que a metodologia avaliada agrupa de forma eficiente o conjunto de dados com base no índice utilizado.

No primeiro experimento foram feitas diferentes variações de sobreposição. Para isto, foi utilizada uma base de dados sintética que consta de 900 amostras, geradas a partir de distribuições normais aleatórias. Esta base de dados é composta de cinco agrupamentos com 180 amostras em cada grupo, onde existe uma variação de desvio padrão no intervalo de $[0-1]$, de modo que é aumentado o grau de sobreposição a medida que a variação fica próxima de 1. Para cada variação de desvio padrão (sobreposição) são apresentados: o conjunto de dados não rotulado como um problema não supervisionado, a identificação de agrupamentos nas imagens de Matriz de Distâncias, como também os agrupamentos obtidos pela metodologia proposta e finalmente a projeção linear dos dados eliminando a sobreposição. Alguns dos resultados gráficos podem ser evidenciados nas Tabelas 2-5. Em adição, a metodologia proposta foi comparada com algoritmos tradicionais como *Fuzzy C-Means (FCM)* e *K-means Clustering*, onde o número $k$ de grupos foi selecionado com o valor máximo de *Silhouette* quando este foi variado de $k = 2, ..., 5$. Os resultados numéricos correspondentes às métricas de avaliação neste experimento com variação do desvio padrão de $dp = 0.1$ até $dp = 0.8$ são mostrados na Tabela 1.

Tabela 1: Resultados Experimento 1

| ***Base de dados*** | ***Silhouette*** | | | ***Precision*** | | | ***Recall*** | | | ***Jaccard*** | | | ***k*** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***(5 gaussian)*** | ***A*** | ***B*** | ***C*** | ***A*** | ***B*** | ***C*** | ***A*** | ***B*** | ***C*** | ***A*** | ***B*** | ***C*** | ***A*** | ***B*** | ***C*** |
| dp=0.1 | **0.8498** | 0.8446 | 0.5912 | **1** | **1** | 0.5541 | **1** | **1** | **1** | **1** | **1** | 0.5541 | **5** | **5** | 3 |
| dp=0.2 | 0.6860 | 0.6893 | **0.6894** | **1** | 0.9845 | 0.9867 | **1** | 0.9845 | 0.9867 | **1** | 0.9695 | 0.9738 | **5** | **5** | **5** |
| dp=0.3 | 0.5204 | 0.5518 | **0.5537** | **1** | 0.4499 | 0.4498 | **1** | 0.9453 | 0.9506 | **1** | 0.4385 | 0.4395 | **5** | 3 | 3 |
| dp=0.4 | 0.3723 | 0.4909 | **0.4923** | **1** | 0.3722 | 0.3734 | **1** | 0.9589 | 0.9642 | **1** | 0.3663 | 0.3683 | **5** | 2 | 2 |
| dp=0.5 | 0.2726 | 0.4691 | **0.4698** | **0.9977** | 0.3630 | 0.3640 | **0.9977** | 0.9320 | 0.9359 | **0.9955** | 0.3536 | 0.3552 | **5** | 2 | 2 |
| dp=0.6 | 0.1686 | **0.4131** | 0.4128 | **0.9955** | 0.3438 | 0.3444 | **0.9955** | 0.8674 | 0.8671 | **0.9911** | 0.3266 | 0.3271 | **5** | 2 | 2 |
| dp=0.7 | 0.1148 | 0.4031 | **0.4055** | **0.7131** | 0.3271 | 0.3257 | **1** | 0.8249 | 0.8271 | **0.7131** | 0.3058 | 0.3049 | **5** | 2 | 2 |
| dp=0.8 | 0.0552 | 0.3727 | **0.3750** | **0.6379** | 0.3614 | 0.3585 | **0.9373** | 0.6184 | 0.6190 | **0.6118** | 0.2955 | 0.2937 | **4** | 3 | 3 |

**A** representa *Cluster-CV*; **B** representa *FCM*; **C** representa *K-means*.

De forma geral, os resultados deste primeiro experimento mostram que a metodologia proposta consegue identificar corretamente o número $k$ de agrupamentos no conjunto de dados, a exceção de $dp = 0.8$ onde foram identificados apenas 4 agrupamentos. Em contraste, os outros métodos tradicionais se mostraram ineficientes na identificação de agrupamentos, conforme foi aumentando o grau de sobreposição. Também, nas Tabelas 2-5 é evidenciado que conforme aumenta o grau de sobreposição no conjunto de dados, é mais difícil para o método *Cluster-CV* encontrar o número de agrupamentos correto através de VC já que a imagem obtida da matriz de distância possui um ruído maior. No en-

Tabela 2: Análise para $dp = 0.2$, com $s_x = s_y = 1$ e $q = 1.9$

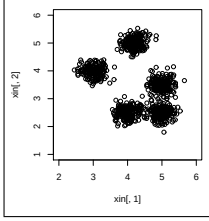
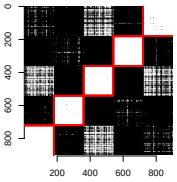
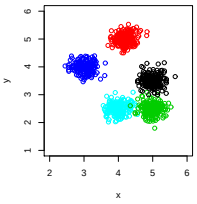
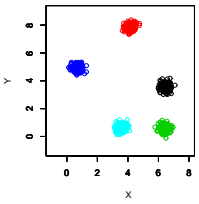
| Não rotulados | Matriz de Distâncias | Metodologia | Projeção |
|---|---|---|---|
|  |  |  |  |

Tabela 3: Análise para $dp = 0.5$, com $s_x = s_y = 1$ e $q = 1.9$

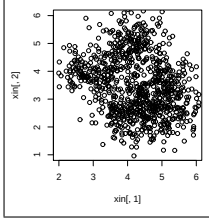
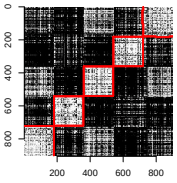
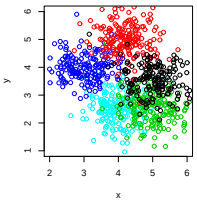
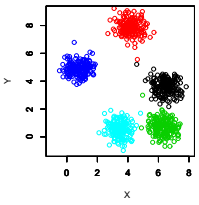
| Não rotulados | Matriz de Distâncias | Metodologia | Projeção |
|---|---|---|---|
|  |  |  |  |

Tabela 4: Análise para $dp = 0.7$, com $s_x = s_y = 0.6$ e $q = 1.9$

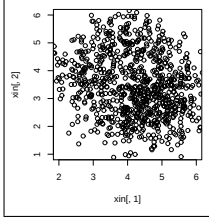
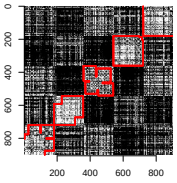
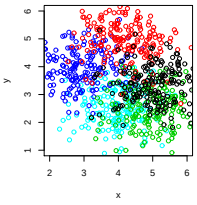
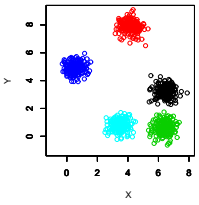
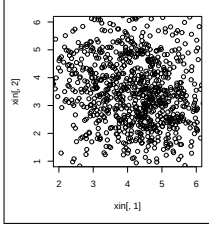
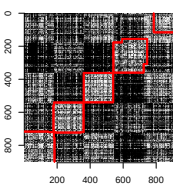
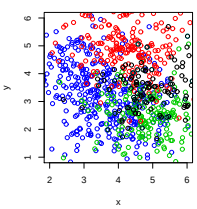
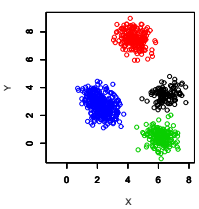
| Não rotulados | Matriz de Distâncias | Metodologia | Projeção |
|---|---|---|---|
|  |  |  |  |

Tabela 5: Análise para $dp = 0.8$, com $s_x = s_y = 0.6$ e $q = 1.9$

| Não rotulados | Matriz de Distâncias | Metodologia | Projeção |
|---|---|---|---|
|  |  |  |  |

tanto os resultados foram satisfatórios mostrando que o *Cluster-CV* é robusto a variações de ruído.

No segundo experimento, a metodologia proposta foi testada sobre seis bases de dados sintéticas, cada uma delas possui 900 amostras, que foram previamente selecionadas do pacote *mlbench* do software *R project*. Cada base de dados foi simulada 30 vezes e foram obtidos os valores médios junto com o desvio padrão para cada uma das métricas de avaliação (*Silhouette*, *Precision*, *Recall* e *Jaccard*). Os resultados deste experimento, apresentados na Tabela 6 mostraram que a metodologia é robusta e fornece resultados bons e confiáveis próximos ou iguais a 1. Isto indica uma ótima qualidade no agrupamento das bases de dados analisadas com características onde os agrupamentos são bem definidos.

Tabela 6: Resultados Experimento 2

| ***Base de dados*** | | ***Silhouette*** | ***Precision*** | ***Recall*** | ***Jaccard*** |
|---|---|---|---|---|---|
| **Cuboids** | **A** | 0.4311±0.0281 | **0.9685±0.0758** | **0.9926±0.0206** | **0.9622±0.0813** |
| | **B** | 0.4704±0.0041 | 0.7536±0.0085 | 0.9485±0.0043 | 0.7240±0.0033 |
| | **C** | **0.4785±0.0052** | 0.7580±0.0025 | 0.9712±0.0134 | 0.7414±0.0102 |
| **Shapes** | **A** | **0.9660±0.0049** | **1±0** | **1±0** | **1±0** |
| | **B** | 0.6977±0.0050 | 1±0 | 1±0 | 1±0 |
| | **C** | 0.6023±0.1010 | 0.8168±0.1930 | 0.9338±0.0702 | 0.7889±0.2225 |
| **Simplex dp=0.25** | **A** | **0.4218±0.0310** | **0.9500±0.1009** | **0.9841±0.0120** | **0.9349±0.0972** |
| | **B** | 0.3176±0.0036 | 0.5766±0.0141 | 0.7762±0.0171 | 0.4946±0.0172 |
| | **C** | 0.3268±0.0059 | 0.5922±0.0158 | 0.8540±0.0244 | 0.5379±0.0214 |
| **Smiley dp1=0.15 dp2=0.15** | **A** | **0.5409±0.0250** | **0.9550±0.0944** | **0.9993±0.0013** | **0.9545±0.0946** |
| | **B** | 0.5033±0.0649 | 0.8015±0.1105 | 0.7341±0.1249 | 0.6336±0.1462 |
| | **C** | 0.5272±0.0464 | 0.8019±0.1536 | 0.7777±0.0881 | 0.6638±0.1523 |
| **Cassini** | **A** | 0.4047±0.0883 | 0.8085±0.1647 | **0.9764±0.0702** | **0.7977±0.1778** |
| | **B** | 0.3844±0.0309 | **0.8800±0.1303** | 0.8639±0.1137 | 0.7902±0.1807 |
| | **C** | **0.4264±0.0211** | 0.7109±0.0620 | 0.7292±0.0410 | 0.5647±0.0686 |
| **Hypercube dp=0.25** | **A** | **0.3486±0.0296** | **0.9289±0.0929** | **0.98859±0.0051** | **0.9186±0.0896** |
| | **B** | 0.2983±0.0107 | 0.4329±0.0071 | 0.7137±0.0142 | 0.3689±0.0089 |
| | **C** | 0.3150±0.0096 | 0.4856±0.0109 | 0.8047±0.0217 | 0.4345±0.0148 |

**A** representa *Cluster-CV*; **B** representa *FCM*; **C** representa *K-means*.

Finalmente, um terceiro experimento foi feito com a finalidade de testar a metodologia proposta quando existe desbalanceamento de dados, ou seja, quando existe um agrupamento com maior número de dados que o outro. Assim, é utilizada uma base de dados com 900 amostras geradas a partir de distribuições normais aleatórias. Esta base de dados possui dois agrupamentos desbalanceados, onde um deles consta de 100 amostras e o outro de 800 amostras. De forma similar ao primeiro experimento, foi variado o grau de sobreposição (*dp* [0.1 - 0.8]) no conjunto de dados desbalanceado e com sobreposição, e seguidamente foram calculadas as métricas de avaliação. Os resultados deste experimento são mostrados na Tabela 7, e adicionalmente, nas Tabelas 8-11 são ilustrados alguns dos resultados gráficos.

Os resultados das métricas de avaliação mostram que a metodologia também é robusta em bases de dados desbalanceadas e com sobreposição; também foi evidenciado que conforme aumenta o grau de sobreposição faz diminuir o valor da métrica de *Silhouette*, mas após da projeção linear dos dados é possível

fornecer melhores resultados de *Silhouette* como foi evidenciado nos resultados (Tabela 7 - *Silhouette* pos-proj).

Tabela 7: Resultados Experimento 3

| ***Base de dados (2 gaussian)*** | ***Silhouette*** | ***Precision*** | ***Recall*** | ***Jaccard*** | ***Silhouette pos-proj.*** |
|---|---|---|---|---|---|
| dp=0.1 | 0.8225 | 1 | 1 | 1 | **0.9415** |
| dp=0.2 | 0.6449 | 0.9975 | 0.9996 | 0.9972 | **0.8871** |
| dp=0.3 | 0.4846 | 1 | 1 | 1 | **0.8297** |
| dp=0.4 | 0.3363 | 0.9975 | 0.9996 | 0.9972 | **0.7726** |
| dp=0.5 | 0.2599 | 0.9014 | 0.9924 | 0.8953 | **0.7337** |
| dp=0.6 | 0.2176 | 0.9996 | 0.9975 | 0.9972 | **0.8569** |
| dp=0.7 | 0.1649 | 0.9950 | 0.9993 | 0.9944 | **0.7893** |
| dp=0.8 | 0.1277 | 0.9830 | 0.9979 | 0.9810 | **0.7818** |

Tabela 8: Análise para $dp = 0.2$, com $s_x = s_y = 1$ e $q = 1.9$

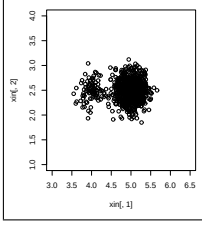
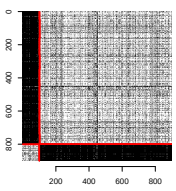
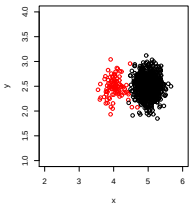
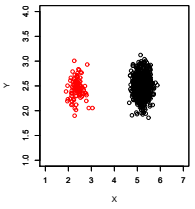
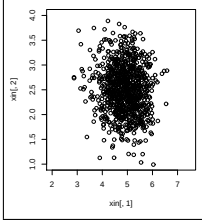
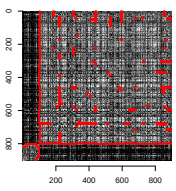
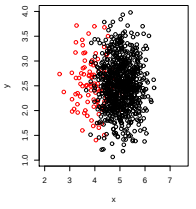
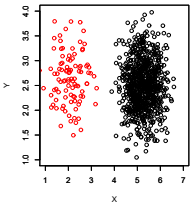
| Não rotulados | Matriz de Distâncias | Metodologia | Projeção |
|---|---|---|---|
|  |  |  |  |

Tabela 9: Análise para $dp = 0.5$, com $s_x = s_y = 1$ e $q = 1.9$

| Não rotulados | Matriz de Distâncias | Metodologia | Projeção |
|---|---|---|---|
|  |  |  |  |

## 5 Conclusão

A metodologia proposta conseguiu atingir os objetivos que foram propostos no desenvolvimento deste trabalho. A análise da matriz de distâncias como imagem, permite extrair informação relevante sobre o número de agrupamentos que possui o conjunto de dados sem precisar de um especialista que forneça o número $k$ de agrupamentos. Assim, a identificação do número de agrupamentos através de Visão Computacional nas imagens mostrou-se robusto para as diferentes variações

Tabela 10: Análise para $dp = 0.7$, com $s_x = s_y = 0.6$ e $q = 2.4$

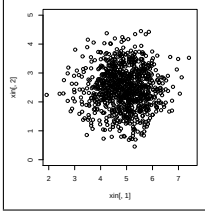
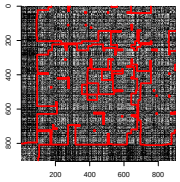
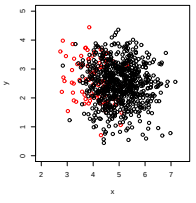
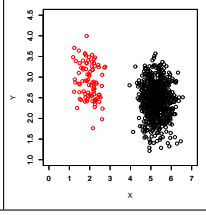
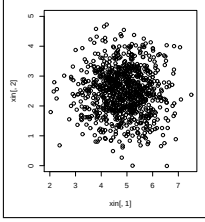
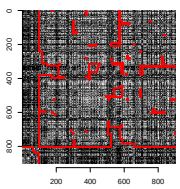
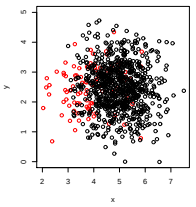
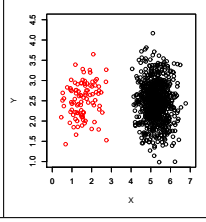
| Não rotulados | Matriz de Distâncias | Metodologia | Projeção |
|---|---|---|---|
|  |  |  |  |

Tabela 11: Análise para $dp = 0.8$, com $s_x = s_y = 0.6$ e $q = 2.4$

| Não rotulados | Matriz de Distâncias | Metodologia | Projeção |
|---|---|---|---|
|  |  |  |  |

de sobreposição, fazendo que em problemas de agrupamento não supervisionado a metodologia forneça bons resultados.

Uma característica significativa da metodologia proposta é que não existe perda de informação ou de amostras na região de sobreposição, porque o método é robusto ao ruído. Desta forma, consegue encontrar os valores pertencentes a cada agrupamento sem ser afetado pela sobreposição. Ao contrário dos outros métodos, que eliminam estas informações com sobreposição nas regiões de fronteira para contornar melhor o problema.

A metodologia proposta demostra que, uma vez identificados o número $k$ de agrupamentos é possível projetar linearmente os dados, de forma que a sobreposição no conjunto original dos dados seja eliminada, reduzindo a distância inter-grupos e aumentado a distância intra-grupos, facilitando assim uma melhor visualização dos dados. Em adição, esta abordagem é independente da inicialização de centroides nos grupos e não precisa de conhecimento prévio dos dados. Ou seja, é não paramétrica dado que não é necessário definir o número $k$ de grupos. Como trabalhos futuros, será testada a metodologia para bases de dados com maior número de dimensões.

## 6 Agradecimentos

O presente trabalho foi realizado com o apoio financeiro da CAPES-Brasil, CNPq e Fapemig.

## Referências

1. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM computing surveys (CSUR) **31**(3) (1999) 264–323

2. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern classification. John Wiley & Sons (2012)
3. Mingoti, S.A.: Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada. Editora UFMG (2005)
4. Goshtasby, A.A.: Image registration: Principles, tools and methods. Springer Science & Business Media (2012)
5. Short, R., Fukunaga, K.: The optimal distance measure for nearest neighbor classification. IEEE transactions on Information Theory **27**(5) (1981) 622–627
6. Perlibakas, V.: Distance measures for pca-based face recognition. Pattern Recognition Letters **25**(6) (2004) 711–724
7. Grzegorzewski, P.: Distances between intuitionistic fuzzy sets and/or interval-valued fuzzy sets based on the hausdorff metric. Fuzzy sets and systems **148**(2) (2004) 319–328
8. De Maesschalck, R., Jouan-Rimbaud, D., Massart, D.L.: The mahalanobis distance. Chemometrics and intelligent laboratory systems **50**(1) (2000) 1–18
9. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. Journal of Machine Learning Research **10**(Feb) (2009) 207–244
10. Bandyopadhyay, S., Saha, S.: Unsupervised classification: similarity measures, classical and metaheuristic approaches, and applications. Springer Science & Business Media (2012)
11. Weiss, Y.: Segmentation using eigenvectors: a unifying view. In: Computer vision, 1999. The proceedings of the seventh IEEE international conference on. Volume 2., IEEE (1999) 975–982
12. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: Advances in neural information processing systems. (2002) 849–856
13. Silvestre, L.J., Lemos, A.P., Braga, J.P., de Pádua Braga, A.: Parameter-free regularization in extreme learning machines with affinity matrices. In: ESANN. (2014)
14. Zhang, S., Wang, R.S., Zhang, X.S.: Identification of overlapping community structure in complex networks using fuzzy c-means clustering. Physica A: Statistical Mechanics and its Applications **374**(1) (2007) 483–490
15. Pérez-Suárez, A., Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A., Medina-Pagola, J.E.: Oclustr: A new graph-based algorithm for overlapping clustering. Neurocomputing **121** (2013) 234–247
16. Shen, J., Chang, S.I., Lee, E.S., Deng, Y., Brown, S.J.: Determination of cluster number in clustering microarray data. Applied Mathematics and Computation **169**(2) (2005) 1172–1185
17. Forsyth, D., Ponce, J.: Computer vision: a modern approach. Upper Saddle River, NJ; London: Prentice Hall (2011)
18. Buades, A., Coll, B., Morel, J.M.: A non-local algorithm for image denoising. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Volume 2., IEEE (2005) 60–65
19. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: Cluster validity methods: part i. ACM Sigmod Record **31**(2) (2002) 40–45
20. Rendón, E., Abundez, I.M., Gutierrez, C., Zagal, S.D., Arizmendi, A., Quiroz, E.M., Arzate, H.E.: A comparison of internal and external cluster validation indexes. In: Proceedings of the 2011 American Conference, San Francisco, CA, USA. Volume 29. (2011)