Universidade Federal de São Carlos

Processo FAPESP: 2024/05031-6

Agrupamento de Dados Baseado em Grafos com Árvores Geradoras Mínimas e Distância de Jensen-Shannon

Bolsista: Matheus dos Santos Sousa

Orientador: Alexandre Luis Magalhães Levada

São Carlos - SP, Brasil

Conteúdo

1 Introdução **2**

2 Fundamentação Teórica **3**
2.1 K-médias 3
2.2 HDBSCAN 4
2.3 Algoritmo de Kruskal 6
2.4 Algoritmo de Agrupamento Divisivo Baseado em MST 7
2.5 Divergência de Jensen-Shannon 7

3 Metodologia Aplicada **8**

4 Resultados Obtidos **10**

5 Conclusão **12**

A Tabelas das Métricas de Avaliação **13**

1 Introdução

Nos últimos anos, o agrupamento de dados ganhou um grande destaque devido à sua ampla gama de áreas de aplicação, especialmente nas áreas de análise da dados, processamento de imagens e reconhecimento de padrão [2, 5, 14]. Trata-se de uma técnica de aprendizado de máquina não supervisionado que consiste em agrupar dados semelhantes, sem que haja rótulos previstos indicando a qual grupo cada amostra pertence.

O algoritmo mais conhecido e utilizado para a realização de agrupamento de dados é o K-médias (*K-means*). Ele é fácil de entender, simples de implementar e apresenta boa eficiência computacional. No entanto, assume que os grupos (*clusters*) possuem forma esférica e tamanhos semelhantes, o que constitui uma limitação importante em contextos com dados de formas ou tamanhos variados [1, 8, 3].

As limitações apresentadas pelo K-médias atualmente foram superadas por métodos mais modenos, como o estado da arte HDBSCAN [4] (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*). Esse é um algoritmo de agrupamento baseado em densidade que constrói uma hierarquia de clusters a partir da árvore geradora mínima das distâncias de alcançabilidade mútua entre os pontos, e em seguida utiliza uma medida de estabilidade para extrair os agrupamentos mais consistentes. Entretanto, apesar de ser um método muito eficiente e robusto quanto a presença de outliers e ruídos, a sua principal limitação está na sensibilidade a múltiplos hiperparâmetros que precisam ser ajustados cautelosamente para cada conjunto de dados a fim de obter resultados satisfatórios.

Um problema recorrente tanto em algoritmos clássicos quanto em algoritmos modernos é o baixo desempenho em conjuntos de dados de alta dimensionalidade (a chamada *maldição da dimensionalidade*) [15]. Em um espaço de alta dimensionalidade, a intuição baseada em distãncias (como a distância euclideana) falha, e os pontos de dados tendem a se tornar "equidistantes"um dos outros. Isso significa que a distinção entre vizinhos próximos e distante se torna menos clara, tornando as métricas de distância menos significativas. Consequentemente, tarefas como agrupamento, classificação e busca de vizinhos mais próximos se tornam muito mais difíceis e ineficientes, pois os algoritmos perdem a capacidade de diferenciar efetivamente as amostras com base em suas distãncias.

Diante das limitações observadas nos algoritmos de agrupamento abordados, especialmente em cenários de alta dimensionalidade, surge a necessidade de investigar abordagens alternativas que contornem esses desafios. Nesse contexto, o uso de métodos baseados em grafos e a aplicação de métricas de distância fundamentadas na Teoria da Informação tem ganhado cada vez mais destaque em problemas relacionados à análise de dados. Com base nisso, este projeto propõe investigar o Algoritmo de Agrupamento Divisivo Baseado em Árvores Geradoras Mínimas (MST, *Minimum Spanning Tree*), utilizando a distância de Jensen-Shannon [12] como medida de dissimilaridade.

Esse algoritmo, além de explorar as relações complexas entre os dados por meio da árvore geradora mínima, recurso também utilizado no HDBSCAN, incorpora também a divergência de Jensen-Shannon, uma medida baseada na entropia relativa, definida no contexto da Teoria da Informação, para a ponderação das arestas dos grafos completos construídos a partir do conjunto de dados.

A escolha da Divergência de Jensen-Shannon (DJS) como métrica para ponderar as arestas dos grafos completos construídos no contexto de agrupamento baseado em Árvores

Geradoras Mínimas (MST) deve-se não apenas à sua robustez em ambientes de alta dimensionalidade, mas também a características intrínsecas que a tornam particularmente vantajosa frente a outras métricas. Em primeiro lugar, a DJS é simétrica e sempre finita, o que a torna uma métrica apropriada para comparar distribuições de probabilidade, característica importante quando os dados são tratados como tais ou convertidos para esse formato. Além disso, a sua base na Teoria da Informação permite capturar relações sutis entre amostras, destacando diferenças que outras métricas, como a Euclidiana ou até mesmo a divergência de Kullback-Leibler, podem não discernir com clareza. Essa propriedade confere à DJS um poder discriminatório superior, especialmente em espaços de alta dimensionalidade, onde muitas métricas sofrem com a "maldição da dimensionalidade". Dessa forma, a DJS proporciona uma estrutura de pesos nas arestas mais sensível e informativa, resultando em grafos mais representativos da estrutura intrínseca dos dados, o que pode levar a agrupamentos mais coerentes e significativos. Com isso, espera-se uma melhora na qualidade dos agrupamentos detectados, especialmente em cenários de alta dimensionalidade.

Para avaliar a eficácia do algoritmo proposto, foram realizados experimentos comparativos com os algoritmos K-médias e o HDBSCAN, além de uma versão do algoritmo divisivo utilizando a distância euclidiana. Todos os métodos foram aplicados aos mesmos conjuntos de dados. Assim, esse projeto busca responder se a integração entre estruturas baseadas em grafos e a distância de Jensen-Shannon é capaz de obter melhores resultados e superar as limitações dos demais métodos aqui citados, especialmente em contextos com dados de alta dimensionalidade.

2 Fundamentação Teórica

2.1 K-médias

O algoritmo K-médias é o mais utilizado entre os métodos de aprendizado não supervisionado. Embora esse método possua diversas variações para diferentes métricas, o foco nesse trabalho será a sua versão que utiliza a distância euclidiana. Resumidamente, o funcionamento dele consiste nos seguintes passos abaixo.

Algorithm 1 Algoritmo K-médias

Dado um conjunto de dados $X = \{x_0, x_1, x_2, \ldots, x_n\}$ com x_i R^d e um inteiro não positivo k, este é o parâmetro que define o número de agrupamentos.

1. Escolha aleatória de k amostras do conjunto.
2. Associar a cada amostra x_i ao agrupamento mais próximo (centróide mais próximo), de acordo com a distância euclidiana.
3. Atualizar os centros dos agrupamentos (média dos pontos do agrupamento)

$$\mu_j^{(t+1)} = \frac{1}{N_k} \sum_{x_i \in \omega_j} x_i$$

   para j = 1, 2, 3, ... , c.

4. Se não houverem mudanças nos rótulos (centros não mudaram), o algoritmo convergiu. Senão, retorne ao passo 2.

O objetivo desse algoritmo é particionar as n amostras do conjunto de dados X em $k < n$ grupos $S = \{s_1, s_2, ..., s_n\}$ para minimizar o espalhamento intra-cluster, ou seja, encontrar uma partição ótima de modo que satisfaça a soma dos quadrados dentro do cluster (within-cluster sum of the squares - *WCSS*) [7]

$$WCSS = \sum_{i=1}^{k} \sum_{x \in S_i} (x - \mu_i)^2$$

onde μ_i é o centróide (ponto médio) da partição s_i Nota-se, pela formulação matemática, deseja-se que os centros dos grupos estejam o mais próximo possível dos dados.

As principais limitações do K-médias a serem destacadas são a restrição a agrupamentos de forma esférica e tamanhos semelhantes, o que torna o algoritmo "cego" para agrupamentos não circularmente simétricos, e o fenômeno da concentração em meio a dados de alta dimensionalidade devido ao uso da distância euclidiana.

2.2 HDBSCAN

O algoritmo Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) é amplamente reconhecido como um dos métodos não supervisionados mais sofisticados e eficientes para a tarefa de agrupamento de dados. Ele se destaca por sua robustez na presença de outliers e ruídos, além de sua capacidade de identificar agrupamentos com tamanhos variados, densidades não uniformes e formatos arbitrários. A seguir, apresenta-se um resumo do funcionamento deste algoritmo [4, 10].

Algorithm 2 Algoritmo HDBSCAN

Dado o conjunto de dados X.

1. Definir a distância central de um ponto $core_k(a)$, ou seja, a distância de um ponto até o k-ésimo vizinho mais próximo

2. Construir um grafo ponderado pela distância de alcançabilidade mútua definida por

$$d_{mreach-k}(a,b) = max\{core_k(a),\, core_k(b),\, d(a,b)\}$$

   , onde $d(a,b)$ é a distância métrica original entre a e b.

3. Obter a MST do grafo ponderado pela distância de alcançabilidade mútua.

4. Obter uma hierarquia de clusters de forma divisiva, ou seja, inicia-se com um único cluster contendo todos os pontos e a cada iteração a i-ésima maior aresta é removida até não existir componentes conectados.

5. Condensar a árvore de hierarquia de clusters, ou seja, percorrer a hierarquia e a cada divisão verificar se um dos novos clusters criados têm menos pontos do que o tamanho mínimo (um dos parâmetros do HDBSCAN). Se sim, declara-se que são pontos que caem fora de um cluster e guarda-se a distância em que isso aconteceu. Em caso contrário, continua-se a divisão e a verificação.

6. Extrair os clusters da árvore. Para isso usa-se uma medida de afinidade com um cluster definida como $\lambda = \frac{1}{distance}$. λ_{birth} e λ_p referem-se a afinidade quando o cluster foi formado e não foi dividido e quando um determinado caiu fora do cluster, respectivamente. E a afinidade de um cluster é definida pela fórmula $\sum_{p \in cluster} (\lambda_p - \lambda_{birth})$. Inicia-se com todos os nós folhas como selecionados e prossegue-se pela árvore em ordem reversa. Se a soma das estabilidades dos clusters filhos for maior que a do pai, então, definimos a estabilidade do cluster pai como a soma dos filhos. Por outro lado, se a estabilidade do pai for maior que a soma dos seus filhos, então declaramos o pai como selecionado e desmarcamos todos os seus descendentes.

7. Por fim, todos os clusters desejados serão retornados.

O HDBSCAN transforma o espaço dos dados de modo que amostras semelhantes fiquem próximas entre si, e identifica agrupamentos a partir da hierarquia extraída da árvore geradora mínima (MST) e do nível de afinidade entre os clusters. No entanto, a obtenção de bons resultados depende da correta configuração de diversos hiperparâmetros específicos para cada conjunto de dados. Além disso, assim como outros métodos, o uso da distância euclidiana pode limitar sua eficácia na análise de dados com atributos em altas dimensões.

2.3 Algoritmo de Kruskal

O algoritmo de Kruskal é um algoritmo clássico da área de grafos utilizado para resolver o problema da Árvore Geradora Mínima (Minimum Spanning Tree – MST). O algoritmo foi proposto por Joseph Kruskal em 1956 e é baseado em uma abordagem gananciosa (greedy) [9]. A ideia central do algoritmo é selecionar, de forma iterativa, as arestas de menor peso que não formem ciclos com as arestas já escolhidas. Assim, deseja-se obter a árvore que minimiza a seguinte função:

$$w(T) = \sum_{e \in T} w(e)$$

ou seja, o objetivo é minimizar o somatório dos pesos das arestas que compõem a árvore.

Para auxiliar na adição de arestas de menor custo sem gerar ciclos, o algoritmo de Kruskal utiliza a estrutura Union-Find fazendo uso das primitivas:

1. Make_Set(v): cria uma árvore contendo um único vértice (raiz).
2. Find_Set(v): retorna qual é a árvore que o vértice v pertence.
3. Union(u, v): faz a fusão das raízes das árvores de u e de v, criando uma única árvore.

A seguir, o pseudocódigo do algoritmo de Kruskal é apresentado.

Algorithm 3 Algoritmo de Kruskal

Dado um grafo convexo, ponderado e não direcionado $G = (V, E, w)$.

1. Crie o conjunto vazio de arestas T.
2. Para cada vértice v pertencente a V, faça Make_Set(v).
3. Criar uma lista de arestas $e \in E$ e ordenar de forma crescente de acordo com o peso w.
4. Para cada aresta $e = (u, v)$ da lista alternada, verificar se Find_Set(u) != Find_Set(v). Se for verdadeiro, adicione e em T e faça Union(u, v).
5. T contém a MST do grafo G.

A complexidade do algoritmo de Kruskal é $O(m * log(n))$, o que é considerada eficiente do ponto de vista computacional, por ser logarítmica em relação ao número de vértices. Por esse motivo, o algoritmo será utilizado como etapa inicial no Algoritmo de Agrupamento Divisivo Baseado em MST, a fim de extrair a árvore geradora mínima do grafo construído a partir dos dados, antes da realização do agrupamento.

2.4 Algoritmo de Agrupamento Divisivo Baseado em MST

Este é um método de agrupamento não supervisionado relativamente simples e de fácil compreensão, assim como o K-médias. Além disso, ele também se beneficia de relações complexas extraídas por meio da árvore geradora mínima (MST), recurso igualmente explorado por algoritmos mais sofisticados como o HDBSCAN. O funcionamento do método pode ser descrito pelos passos apresentados a seguir.

Algorithm 4 Algoritmo de Agrupamento Divisivo Baseado em MST

Dado um conjunto de dados $X = \{x_0, x_1, x_2, ..., x_n\}$ com $x \in R^d$ representado por um grafo ponderado não-direcionado $G = (V, E, w)$; $V = \{x_0, x_1, ..., x_n\}$, $E = \{\{u, v\}, u < v\}$, $W(\{u, v\}) = ||x_u - x_v||$ e um inteiro não-negativo k.

1. Obter a árvore geradora mínima $T = MST(G) = (V, E', W')$.
2. Obter $\{\{x_0, x_1, ..., x_n\}\}$, a primeira partição inicial representadas por todos os pontos.
3. Para $i = 1, ..., k$ faça:
   (a) Dividir o cluster que contenha os u-ésimo e v-ésimo pontos (de modo que eles não pertençam mais ao mesmo componente), onde $\{u, v\} \in W$ é um aresta da MST com o i-ésimo maior tamanho.
4. Retorna a atual k-partição como resultado.

O objetivo desse algoritmo é particionar, por meio da árvore geradora mínima do grafo dado as n amostras em $k < n$ grupos $S = \{s_0, s_1, s_2, ..., s_k\}$ de modo a maximizar a soma das $k-1$ arestas omitidas durante a obtenção de clusters de forma iterativa. É importante ressaltar que o valor k é definido a priori.

Do ponto de vista computacional, observa-se que o maior custo do algoritmo está concentrado na construção da MST. Como mencionado anteriormente, utilizando o algoritmo de Kruskal, essa etapa apresenta complexidade $O(n * log(n))$. Neste trabalho, a fim de contornar as limitações impostas pelo uso da distância euclidiana em espaços de alta dimensionalidade, será empregada a divergência de Jensen–Shannon como medida alternativa para ponderar as arestas do grafo.

2.5 Divergência de Jensen-Shannon

Conforme discutido anteriormente, o fenômeno conhecido como mal da dimensionalidade afeta negativamente diversos modelos que utilizam a distância euclidiana como métrica, especialmente em conjuntos de dados de alta dimensionalidade, comprometendo, assim, o desempenho desses modelos.

Uma alternativa para contornar esse problema é utilizar métricas de distância não euclidianas, como aquelas fundamentadas na Teoria da Informação. Essas medidas permitem calcular a dissimilaridade entre distribuições de probabilidade, oferecendo maior

robustez em contextos de alta dimensionalidade. Dentre elas, destaca-se a divergência de Jensen-Shannon (DJS), uma métrica simétrica e finita baseada na entropia relativa, utilizada para quantifica a similaridade entre duas distribuições de probabilidade [12, 13]. Primeiramente, define-se a divergência de Kullback-Leibler (KL) entre duas distribuições de probabilidade P e Q como a entropia relativa:

$$D_{KL}(P, Q) = H(P, Q) - H(P)$$

onde $H(P)$ é a entropia de Shannon de P e $H(P, Q)$ é a entropia de Shannon cruzada entre P e Q, definidas como:

$$H(P) = \sum_{i=1}^{n} p_i \log(p_i)$$

$$H(P, Q) = \sum_{i=1}^{n} p_i \log(q_i)$$

O problema com a divergência de KL é que ela não é uma métrica, pois pode-se mostrar que $D_{KL}(P, Q) \neq D_{KL}(Q, P)$. Uma forma de tornar a divergência de KL simétrica é a partir da divergência de Jensen-Shannon, dada por:

$$D_{JS}(P, Q) = \frac{1}{2} D_{KL}(P, M) + \frac{1}{2} D_{KL}(M, Q)$$

onde $M = \frac{1}{2}(P + Q)$ é a média das distribuições P e Q. Outra forma de expressar a divergência de Jensen-Shannon é:

$$D_{JS}(P, Q) = H(M) - \frac{H(P) + H(Q)}{2}$$

Nota-se que, quando duas distribuições são totalmente sobrepostas, ou seja, quando $P = Q$, a divergência de Jensen-Shannon é igual a zero. Por outro lado, quanto menor a sobreposição entre P e Q, maior será o valor da divergência, refletindo uma maior dissimilaridade entre as distribuições.

O fato de a divergência de Jensen-Shannon (DJS) ser mais discriminante do que a distância Euclidiana, isto é, possuir maior capacidade de distinguir amostras pertencentes a diferentes grupos, justifica sua vantagem na análise de conjuntos de dados em alta dimensionalidade.

3 Metodologia Aplicada

Neste projeto, adotou-se uma abordagem comparativa entre diferentes algoritmos de agrupamento, com o objetivo de avaliar a eficácia do método divisivo baseado em árvores geradoras mínimas (MST) utilizando a divergência de Jensen-Shannon. Essa abordagem foi comparada com três métodos: o algoritmo K-médias, o HDBSCAN e uma versão da mesma técnica divisiva com MST, porém utilizando a distância Euclidiana. A proposta visa investigar se a utilização da divergência de Jensen-Shannon, uma métrica mais

informativa e sensível a diferenças em distribuições de probabilidade, é capaz de superar as limitações enfrentadas pelos demais algoritmos, especialmente em cenários de alta dimensionalidade, onde a distância Euclidiana tende a perder poder discriminativo.

A linguagem de programação Python foi utilizada para a implementação do Algoritmo de Kruskal, do algoritmo de agrupamento divisivo baseado em MST com as distâncias Euclidiana e de Jensen-Shannon, além do algoritmo K-médias. Devido à complexidade da implementação do HDBSCAN, optou-se por utilizar a versão disponível na biblioteca Scikit-Learn. Para dar suporte ao desenvolvimento, foram empregadas bibliotecas amplamente utilizadas na comunidade científica, como NumPy, SciPy, Pandas e NetworkX, as quais facilitaram o processamento dos dados, operações matriciais e manipulação de grafos.

Os conjuntos de dados utilizados neste estudo foram obtidos do repositório OpenML e estão resumidos na Tabela 1. A maioria dos conjuntos rotulados com o prefixo "AP" consiste em dados de expressão gênica derivados do projeto GEMLeR, com um alto número de atributos (10.935), refletindo a natureza de dados de microarranjos. Esses conjuntos envolvem tarefas de classificação binária entre diferentes tipos de tecidos ou órgãos. Além disso, foram incluídos conjuntos variados, como MNIST_784, Fashion-MNIST e mfeat-factors, com o objetivo de avaliar o desempenho dos algoritmos em diferentes contextos, desde dados visuais até dados biomédicos e de voz. A tabela apresenta, para cada conjunto, o número de amostras, o número de atributos e o número de classes

Tabela 1: Número de amostras, atributos e classes dos conjuntos de dados utilizados.

ID	Conjunto de Dados	Amostras	Atributos	Classes
1	Iris	150	4	3
2	Wine	178	13	3
3	Digits	1797	64	10
4	AP_Lung_Kidney	386	10935	2
5	AP_Colon_Kidney	546	10935	2
6	AP_Colon_Uterus	410	10935	2
7	AP_Breast_Uterus	468	10935	2
8	AP_Breast_Colon	630	10935	2
9	AP_Colon_Omentum	363	10935	2
10	AP_Breast_Omentum	421	10935	2
11	AP_Omentum_Kidney	421	10935	2
12	AP_Omentum_Uterus	201	10935	2
13	AP_Colon_Prostate	355	10935	2
14	AP_Lung_Uterus	250	10935	2
15	AP_Breast_Prostate	413	10935	2
16	parkinson-speech-uc	756	753	2
17	semeion	1593	256	10
18	mfeat-factors	2000	216	10
19	Olivetti_Faces	400	4096	40
20	oh5.wc	918	3012	10
21	dbworld-bodies	64	4702	2
22	oh15.wc	913	3100	10
23	tr31.wc	927	10128	7
24	BurkittLymphoma	220	22283	3
25	ovarianTumour	283	54621	3
26	hepatitisC	283	54621	3
27	micro-mass	571	1300	20
28	scene	2407	299	2
29	musk	6568	167	2
30	Speech	3686	400	2
31	MNIST_784	3500	784	10
32	Fashion-MNIST	3500	784	10

Para avaliar o desempenho dos algoritmos de agrupamento, foi utilizada uma variedade de métricas internas e externas, a fim de obter uma análise abrangente da qualidade dos agrupamentos gerados. Entre as métricas externas, que dependem de rótulos verdadeiros para comparação, utilizou-se: Rand Index, Adjusted Rand Index, Mutual Information Score, Adjusted Mutual Information, Fowlkes-Mallows, Homogeneity, Completeness e V-Measure. Já as métricas internas, que avaliam a coerência dos agrupamentos com base apenas nos dados e agrupamentos gerados, incluem: Silhouette Coefficient, Calinski-Harabasz Score e Davies-Bouldin Score. Esse conjunto de métricas permite uma avaliação equilibrada entre fidelidade aos rótulos reais e coesão/separação dos grupos formados, permitindo uma melhor comparação entre os métodos testados.

Vale destacar que, devido à aleatoriedade na escolha inicial dos centróides, o algoritmo K-médias pode gerar diferentes resultados a cada execução. Por esse motivo, ele foi executado 20 vezes para cada conjunto de dados, sendo considerada a média dos valores obtidos em cada métrica de avaliação. Em contrapartida, os demais algoritmos, por serem determinísticos e produzirem resultados consistentes para os mesmos dados de entrada, foram executados apenas uma vez. Ademais, considerando a complexidade envolvida no ajuste dos hiperparâmetros do HDBSCAN, optou-se por utilizar uma configuração fixa para todos os conjuntos de dados: `HDBSCAN(min_samples=2, min_cluster_size=2, metric='euclidean', cluster_selection_method='eom')`.

Por fim, foi realizada uma análise estatística para verificar se as diferenças de desempenho entre os algoritmos são estatisticamente significativas. Para isso, aplicou-se o teste de hipótese de Friedman [6] para cada métrica de avaliação. Caso o valor-p obtido no teste de Friedman fosse inferior a 0,05, considerava-se que havia diferença significativa entre os algoritmos, e então procedia-se com o teste de post hoc de Nemenyi[11]. Como base na matriz de valores p desse teste, eram identificados os algoritmos cujos resultados diferiam significamente do algoritmo divisivo baseado em MST com divergência de JS, ou seja, pares com $p < 0,05$. Nesses casos, comparavam-se as medianas dos resultados obtidos para determinar qual algoritmo apresentou melhor desempenho em cada métrica.

4 Resultados Obtidos

Devido à quantidade e ao tamanho das tabelas com os resultados brutos dos quatro algoritmos analisados, optou-se por apresentá-las na seção de Apêndice (Apêndice A). Logo a seguir, estão os resultados do teste de hipótese de Friedman.

Tabela 2: Valores-p do teste de hipótese de Friedman para cada métrica

Nome	**Valor-p**
Rand Index	0,7345
Adj. Rand Index	$4,8775 \times 10^{-15}$
Adj. Mutual Info	$4,2950 \times 10^{-10}$
Mutual Info	$3,6760 \times 10^{-11}$
Fowlkes Mallows Index	$2,5610 \times 10^{-6}$
Homogeneity Score	$3,6760 \times 10^{-11}$
Completeness Score	0,5082
V Measure	$8,7036 \times 10^{-10}$
Silhouette	$7,9882 \times 10^{-5}$
Calinski	$2,3066 \times 10^{-14}$
Davies	$8,7734 \times 10^{-15}$

Na tabela 2 observou-se que, para as métricas Rand Index e Completeness Score, o valor-p foi superior a 0,05, indicando ausência de diferença estatística significativa entre os algoritmos, ou seja, seus desempenhos podem ser considerados equivalentes nessas métricas. Para as demais métricas de avaliação, foi aplicado o teste post hoc de Nemenyi, com o objetivo de comparar o desempenho de cada par de algoritmos. A análise concentrou-se especificamente nos pares que envolvem o Algoritmo Divisivo Baseado em árvore geradora mínima (MST) com Jensen-Shannon. Sempre que o valor-p obtido foi inferior a 0,05, indicando diferença estatística significativa, foi realizada uma comparação adicional entre as medianas das métricas de avaliação para determinar qual método apresentou desempenho superior. Os resultados dessas comparações estão apresentados a seguir:

Tabela 3: Comparações entre Algoritmo Divisivo com JS e demais algoritmos (valor-$p < 0,05$ para o teste de Nemenyi)

Métrica	**Algoritmo comparado**	**Mediana superior**	**Desempenho melhor**
Adjusted Rand Index	K-médias	K-médias	K-médias
Adjusted Mutual Info	K-médias / HDBSCAN	K-médias / HDBSCAN	K-médias / HDBSCAN
Mutual Info	K-médias / HDBSCAN	K-médias / HDBSCAN	K-médias / HDBSCAN
Fowlkes-Mallows	K-médias / HDBSCAN	Divisivo com JS	Divisivo com JS
Homogeneity	K-médias / HDBSCAN	K-médias / HDBSCAN	K-médias / HDBSCAN
V-Measure	K-médias / HDBSCAN	K-médias / HDBSCAN	K-médias / HDBSCAN
Silhouette	Divisivo com dist. eucl.	Divisivo com dist. eucl.	Divisivo com dist. eucl.
Calinski-Harabasz	K-médias / HDBSCAN	K-médias / HDBSCAN	K-médias / HDBSCAN
Davies-Bouldin	K-médias / HDBSCAN	K-médias / HDBSCAN	Divisivo com JS

Nota-se, por meio da Tabela 3, que o algoritmo divisivo com divergência de Jensen-Shannon apresentou desempenho inferior nas métricas Adjusted Mutual Information, Mutual Information, Fowlkes-Mallows, Homogeneity, V-Measure e Calinski-Harabasz, quando comparado ao K-médias e ao HDBSCAN, sendo, contudo, equivalente ao algoritmo divisivo com distância Euclidiana.

Em relação à métrica Adjusted Rand Index, o desempenho do algoritmo divisivo com Jensen-Shannon foi inferior ao do K-médias, mas equivalente ao HDBSCAN e ao divisivo com distância Euclidiana.

Já na métrica Silhouette, o divisivo com Jensen-Shannon obteve resultados inferiores aos do divisivo com distância Euclidiana, mas semelhantes aos apresentados pelo K-médias e pelo HDBSCAN.

Por outro lado, nas métricas Fowlkes-Mallows e Davies-Bouldin, o algoritmo divisivo com divergência de Jensen-Shannon apresentou desempenho superior aos demais. No caso da métrica Fowlkes-Mallows, que avalia a concordância entre os agrupamentos gerados e as classes reais, o divisivo com JS obteve os melhores resultados, indicando maior precisão na formação dos clusters. De forma semelhante, na métrica Davies-Bouldin, que mede a separação e compacidade dos agrupamentos (sendo valores menores melhores), o divisivo com JS também foi o que apresentou o melhor desempenho, evidenciando sua capacidade de gerar clusters bem definidos e internamente coesos.

5 Conclusão

Este projeto propôs e avaliou algoritmo de agrupamento divisivo baseado em árvore geradora mínima (MST), utilizando a divergência de Jensen-Shannon como medida de dissimilaridade entre as amostras. A abordagem foi criada para lidar com conjuntos de dados em alta dimensionalidade, um cenário bastante desafiador tanto para algoritmos clássicos quanto modernos, no qual enfrentam limitaçôes quanto à definição de distâncias e à separabilidade dos grupos.

A análise comparativa, fundamentada em testes estatísticos robustos, demonstrou que o algoritmo proposto apresenta desempenho competitvo em relação a métodos consagrados como o K-médias e o HDBSCAN. Apesar de seu desempenho inferior em algumas métricas, como Adjusted Mutual Information, Mutual Information, Homogeneity, V-measure e Calinski-Harabasz, o método obteve resultados superiores nas métricas Fowlkes-Mallows e Davies-Bouldin. Tais métricas avaliam, respectivamente, a concordância entre os agrupamentos e as classes reais, e a coesão interna e separação dos clusters, sendo altamente relevantes na qualidade do agrupamento. Esses resultados sugerem que o algoritmo divisivo baseado em MST com divergência de Jensen-Shannon possui maior aptidão para identificar agrupamentos bem definidos, especialmente em contextos com elevada dimensionalidade, nos quais a estrutura dos dados pode ser sutil ou sobreposta.

Além disso, o desempenho equivalente aos métodos HDBSCAN e K-médias em métricas como Rand Index, Completeness Score e Silhouette Coefficient indica que o algoritmo mantém estabilidade e coerência em diferentes perspectivas de avaliação, mesmo diante da complexidade imposta pela alta dimensionalidade dos dados.

Assim, conclui-se que a utilização da distância de Jensen-Shannon em uma estratégia divisiva baseada em MST configura-se como uma alternativa viável e promissora para tarefas de agrupamento em ambientes de alta dimensionalidade. Diante da capacidade dessa métrica em capturar relações relevantes entre os dados, como trabalhos futuros, propõe-se a aplicação da divergência de JS em outros algoritmos de agrupamento como o próprio K-médias e o HDBSCAN para investigar possíveis novas perspectivas sobre a sua eficácia em detectar estruturas complexas em cenários de alta dimensionalidade. Sugere-se também o uso dessa métrica em tarefas de redução de dimensionalidade, especialmente em técnicas não lineares baseadas em aprendizado de máquina como autoencoders, t-SNE e UMAP. Nesse contexto, a métrica pode ser utilizada como critério de similaridade entre amostras, contribuindo para projeções que preservem relações probabilísticas e semâticas entre os dados originais.

A Tabelas das Métricas de Avaliação

Tabela 4: Resultados do K-médias para cada conjunto de dados (identificados pelo ID da Tabela 1)

ID	Rand Index	Adj. Rand Index	Mutual Info	Adj. Mut. Info	Fowlkes Mall.	Homog. Score	Complet. Score	V Meas.	Silhou. Coeff.	Calinski Har.	Davies Boul.
1	0,853	0,679	0,783	0,722	0,793	0,712	0,741	0,726	0,547	18,967	0,713
2	0,713	0,367	0,459	0,421	0,586	0,422	0,433	0,427	0,568	548,853	0,537
3	0,927	0,621	1,652	0,620	0,664	0,717	0,740	0,728	0,177	162,811	1,885
4	0,585	0,165	0,099	0,149	0,619	0,158	0,145	0,151	0,149	56,779	2,343
5	0,539	0,078	0,048	0,069	0,546	0,070	0,071	0,071	0,165	103,838	2,193
6	0,501	0,004	0,003	0,003	0,538	0,005	0,005	0,005	0,196	119,746	1,761
7	0,499	−0,002	0,000	−0,001	0,552	0,000	0,000	0,000	0,176	110,055	1,980
8	0,508	0,016	0,007	0,010	0,512	0,011	0,011	0,011	0,180	151,016	1,966
9	0,500	−0,001	0,006	0,008	0,577	0,013	0,009	0,011	0,203	107,495	1,757
10	0,502	0,000	0,000	−0,001	0,595	0,000	0,000	0,000	0,178	99,819	1,965
11	0,510	−0,015	0,005	0,007	0,592	0,010	0,012	0,010	0,181	56,625	2,140
12	0,501	0,000	0,001	−0,001	0,522	0,002	0,002	0,002	0,200	56,948	1,751
13	0,619	0,234	0,116	0,227	0,685	0,237	0,223	0,229	0,181	83,057	1,868
14	0,498	−0,003	0,000	−0,002	0,517	0,000	0,000	0,000	0,183	46,322	2,070
15	0,525	0,041	0,022	0,046	0,623	0,048	0,048	0,048	0,159	82,102	2,117
16	0,614	0,176	0,042	0,074	0,691	0,075	0,076	0,075	0,631	1016,728	0,600
17	0,723	0,091	0,386	0,188	0,244	0,167	0,234	0,195	0,005	22,492	5,239
18	0,898	0,480	1,388	0,611	0,539	0,603	0,626	0,614	0,226	372,290	1,472
19	0,964	0,366	2,651	0,580	0,392	0,718	0,764	0,740	0,113	13,207	1,796
20	0,569	0,026	0,306	0,154	0,242	0,136	0,245	0,173	−0,006	11,116	2,891
21	0,507	0,020	0,022	0,017	0,692	0,032	0,140	0,044	0,036	1,334	1,154
22	0,577	0,057	0,347	0,180	0,277	0,155	0,284	0,199	0,018	14,259	2,814
23	0,367	0,033	0,091	0,075	0,472	0,059	0,197	0,090	0,543	1603,898	1,414
24	0,666	0,305	0,387	0,377	0,581	0,399	0,370	0,383	0,101	23,877	2,535
25	0,497	0,086	0,106	0,137	0,594	0,219	0,110	0,146	0,059	14,867	3,490
26	0,477	0,068	1,103	0,130	0,572	0,212	0,102	0,138	0,055	15,083	3,622
27	0,767	0,088	1,131	0,375	0,210	0,388	0,531	0,448	0,102	29,716	1,594
28	0,501	−0,066	0,057	0,106	0,614	0,122	0,094	0,106	0,163	67,974	2,348
29	0,509	−0,019	0,012	0,023	0,623	0,029	0,019	0,023	0,287	2831,158	1,394
30	0,527	0,001	0,000	0,003	0,713	0,007	0,024	0,003	0,015	5,139	8,313
31	0,883	0,379	1,164	0,509	0,444	0,506	0,516	0,511	0,063	135,387	2,820
32	0,878	0,369	1,198	0,527	0,437	0,520	0,540	0,530	0,142	434,335	1,962

Tabela 5: Resultados do HDBSCAN para cada conjunto de dados (identificados pelo ID da Tabela 1)

ID	Rand Index	Adj. Rand Index	Mutual Info	Adj. Mut. Info	Fowlkes Mall.	Homog. Score	Complet. Score	V Meas.	Silhou. Coeff.	Calinski Har.	Davies Boul.
1	0,776	0,568	0,637	0,732	0,771	0,579	1,000	0,734	0,687	502,822	0,383
2	0,663	0,297	0,440	0,381	0,569	0,405	0,384	0,394	0,418	147,066	2,688
3	0,896	0,483	1,737	0,727	0,544	0,754	0,714	0,734	0,072	71,449	1,954
4	0,558	0,107	0,044	0,062	0,598	0,070	0,062	0,066	0,047	13,399	3,223
5	0,503	0,005	0,010	0,012	0,574	0,014	0,017	0,015	0,063	20,498	2,813
6	0,531	−0,008	0,060	0,102	0,644	0,098	0,117	0,106	−0,070	7,192	1,975
7	0,602	0,047	0,041	0,089	0,743	0,070	0,141	0,094	−0,104	8,823	1,924
8	0,508	0,011	0,031	0,062	0,679	0,046	0,126	0,067	−0,149	7,093	1,925
9	0,606	−0,065	0,014	0,027	0,755	0,026	0,054	0,035	−0,148	6,328	1,954
10	0,564	0,001	0,003	0,000	0,679	0,005	0,005	0,005	−0,104	15,629	2,496
11	0,547	0,054	0,022	0,027	0,630	0,041	0,030	0,035	0,086	13,223	2,491
12	0,491	−0,017	0,032	0,037	0,501	0,049	0,040	0,044	−0,022	15,078	2,770
13	0,606	0,271	0,294	0,394	0,660	0,597	0,297	0,396	0,093	41,495	2,309
14	0,515	0,030	0,043	0,054	0,535	0,062	0,058	0,060	−0,029	15,873	2,356
15	0,667	0,310	0,238	0,377	0,744	0,527	0,296	0,379	0,001	24,475	2,680
16	0,385	0,000	0,102	0,025	0,123	0,181	0,026	0,045	0,422	26,773	3,731
17	0,767	0,203	1,220	0,542	0,346	0,530	0,590	0,558	−0,028	14,390	2,670
18	0,811	0,263	1,278	0,561	0,381	0,555	0,595	0,575	−0,056	59,363	1,639
19	0,841	0,126	2,271	0,567	0,249	0,616	0,838	0,710	0,035	9,350	1,959
20	0,235	−0,001	0,040	0,019	0,305	0,018	0,124	0,031	0,184	4,893	4,094
21	0,580	0,160	0,108	0,104	0,567	0,157	0,122	0,138	0,092	2,994	2,798
22	0,418	0,013	0,171	0,068	0,282	0,077	0,216	0,113	−0,137	2,624	1,999
23	0,515	−0,014	0,422	0,113	0,327	0,273	0,208	0,236	−0,284	0,365	2,065
24	0,448	0,020	0,017	0,011	0,631	0,017	0,081	0,029	0,123	2,161	3,797
25	0,695	0,148	0,057	0,096	0,802	0,116	0,120	0,118	0,074	3,525	2,919
26	0,695	0,148	0,057	0,096	0,802	0,116	0,120	0,118	0,074	3,525	2,919
27	0,712	0,004	1,308	0,188	0,127	0,449	0,450	0,450	−0,126	2,567	1,996
28	0,550	−0,018	0,085	0,047	0,667	0,181	0,057	0,086	−0,264	2,093	1,607
29	0,267	0,001	0,361	0,056	0,092	0,840	0,051	0,096	0,315	17,621	1,138
30	0,087	−0,006	0,041	−0,001	0,239	0,488	0,007	0,014	0,104	2,453	1,786
31	0,722	0,003	1,207	0,207	0,152	0,525	0,299	0,381	−0,113	3,245	1,546
32	0,627	0,008	0,897	0,175	0,194	0,390	0,285	0,330	−0,249	3,923	1,421

Tabela 6: Resultados do Algoritmo Divisivo Baseado em MST com distância euclidiana para cada conjunto de dados (identificados pelo ID da Tabela 1)

ID	Rand Index	Adj. Rand Index	Mutual Info	Adj. Mut. Info	Fowlkes Mall.	Homog. Score	Complet. Score	V Meas.	Silhou. Coeff.	Calinski Har.	Davies Boul.
1	0,777	0,564	0,646	0,713	0,764	0,588	0,920	0,717	0,512	277,995	0,447
2	0,363	0,005	0,042	0,565	0,565	0,035	0,236	0,062	0,488	24,420	0,308
3	0,108	0,000	0,012	0,000	0,314	0,005	0,274	0,010	−0,134	1,246	0,924
4	0,557	−0,003	0,001	−0,002	0,745	0,002	0,057	0,003	0,434	4,862	0,430
5	0,500	0,000	0,001	0,000	0,706	0,002	0,102	0,004	0,643	13,774	0,258
6	0,575	−0,003	0,001	−0,002	0,757	0,001	0,051	0,003	0,454	5,353	0,411
7	0,606	−0,005	0,001	0,000	0,776	0,002	0,048	0,004	0,616	21,870	0,649
8	0,503	−0,001	0,002	0,002	0,707	0,003	0,090	0,005	0,610	20,886	0,657
9	0,662	−0,004	0,001	−0,003	0,812	0,001	0,035	0,002	0,448	5,156	0,418
10	0,694	−0,007	0,001	−0,001	0,832	0,002	0,032	0,004	0,614	21,420	0,652
11	0,643	−0,004	0,001	−0,003	0,801	0,001	0,038	0,003	0,644	13,811	0,256
12	0,527	0,006	0,005	0,004	0,723	0,007	0,153	0,014	0,513	7,443	0,352
13	0,683	−0,004	0,001	−0,003	0,825	0,001	0,032	0,002	0,450	5,263	0,416
14	0,498	0,000	0,003	0,000	0,703	0,004	0,105	0,008	0,564	9,058	0,320
15	0,715	−0,008	0,001	−0,002	0,844	0,002	0,029	0,004	0,613	21,403	0,655
16	0,622	0,005	0,002	0,004	0,788	0,003	0,179	0,006	0,887	97,831	0,074
17	0,110	0,000	0,015	0,001	0,313	0,006	0,281	0,012	−0,038	1,359	0,962
18	0,107	0,000	0,010	0,000	0,314	0,005	0,268	0,009	−0,183	1,983	0,790
19	0,599	0,040	1,463	0,359	0,182	0,397	0,812	0,533	−0,012	5,068	1,149
20	0,126	0,000	0,021	0,000	0,330	0,009	0,274	0,018	0,264	3,644	0,516
21	0,500	0,006	0,013	0,004	0,696	0,018	0,156	0,033	0,314	3,149	0,542
22	0,131	0,000	0,023	0,001	0,337	0,010	0,295	0,020	0,327	4,661	0,454
23	0,255	−0,001	0,009	−0,001	0,495	0,006	0,185	0,012	0,902	1038,246	0,045
24	0,421	−0,010	0,005	−0,008	0,640	0,005	0,085	0,010	0,194	2,554	0,641
25	0,759	0,033	0,010	0,025	0,869	0,020	0,212	0,037	0,158	2,017	0,705
26	0,759	0,033	0,010	0,025	0,869	0,020	0,212	0,037	0,158	2,017	0,705
27	0,114	−0,002	0,092	−0,003	0,229	0,032	0,379	0,059	0,291	7,332	0,365
28	0,705	0,000	0,000	0,000	0,840	0,000	0,022	0,000	0,361	3,779	0,499
29	0,738	−0,001	0,000	0,000	0,859	0,000	0,021	0,001	0,262	3,692	0,984
30	0,966	−0,001	0,000	0,000	0,983	0,000	0,002	0,000	0,129	3,155	0,939
31	0,104	0,000	0,006	0,000	0,316	0,003	0,252	0,005	0,007	1,697	0,782
32	0,104	0,000	0,006	0,000	0,315	0,003	0,252	0,005	0,031	1,803	0,748

Tabela 7: Resultados do Algoritmo Divisivo Baseado em MST com divergência de Jensen-Shannon para cada conjunto de dados (identificados pelo ID da Tabela 1)

ID	Rand Index	Adj. Rand Index	Mutual Info	Adj. Mut. Info	Fowlkes Mall.	Homog. Score	Complet. Score	V Meas.	Silhou. Coeff.	Calinski Har.	Davies Boul.
1	0,772	0,558	0,637	0,716	0,764	0,579	0,951	0,720	0,554	253,062	0,375
2	0,346	0,000	0,013	0,001	0,575	0,012	0,181	0,022	−0,401	0,700	1,194
3	0,108	0,000	0,012	0,000	0,314	0,005	0,272	0,010	−0,150	1,139	0,946
4	0,559	0,003	0,000	−0,004	0,745	0,001	0,010	0,001	0,599	21,070	0,452
5	0,500	0,000	0,001	0,000	0,706	0,002	0,102	0,004	0,643	13,774	0,258
6	0,575	−0,003	0,001	−0,002	0,757	0,001	0,051	0,003	0,454	5,353	0,411
7	0,608	−0,003	0,001	−0,002	0,778	0,001	0,043	0,002	0,316	3,001	0,550
8	0,504	0,001	0,001	0,000	0,709	0,002	0,106	0,004	0,445	5,111	0,421
9	0,662	−0,004	0,001	−0,003	0,812	0,001	0,035	0,002	0,448	5,156	0,418
10	0,703	0,016	0,004	0,012	0,838	0,009	0,242	0,016	0,039	1,178	0,873
11	0,643	−0,004	0,001	−0,003	0,801	0,001	0,038	0,003	0,644	13,811	0,256
12	0,527	0,006	0,005	0,004	0,723	0,007	0,153	0,014	0,055	1,281	0,842
13	0,683	−0,004	0,001	−0,003	0,825	0,001	0,032	0,002	0,450	5,263	0,416
14	0,498	0,000	0,003	0,000	0,703	0,004	0,105	0,008	0,564	9,058	0,320
15	0,718	−0,004	0,000	−0,003	0,846	0,001	0,026	0,002	0,305	2,880	0,562
16	0,619	−0,002	0,000	−0,001	0,786	0,001	0,038	0,001	−0,148	0,857	0,763
17	0,109	0,000	0,013	0,000	0,314	0,006	0,277	0,011	−0,094	0,965	1,029
18	0,108	0,000	0,012	0,001	0,314	0,005	0,273	0,010	−0,193	2,032	0,841
19	0,513	0,024	1,244	0,275	0,163	0,337	0,775	0,470	−0,050	4,668	1,138
20	0,127	0,000	0,023	0,001	0,331	0,010	0,298	0,020	−0,270	0,381	1,783
21	0,494	−0,005	0,010	−0,004	0,692	0,014	0,118	0,025	−0,191	0,346	1,632
22	0,131	0,001	0,023	0,001	0,337	0,010	0,296	0,020	−0,270	0,430	1,727
23	0,261	0,006	0,021	0,013	0,500	0,013	0,358	0,026	−0,531	0,020	2,049
24	0,427	0,002	0,009	0,001	0,646	0,010	0,162	0,018	0,234	2,736	0,607
25	0,759	0,033	0,010	0,025	0,869	0,020	0,212	0,037	0,158	2,017	0,705
26	0,759	0,033	0,010	0,025	0,869	0,020	0,212	0,037	0,158	2,017	0,705
27	0,778	0,252	1,573	0,651	0,424	0,540	0,919	0,680	−0,286	8,442	1,808
28	0,705	0,000	0,000	0,000	0,840	0,000	0,022	0,000	0,252	2,511	0,612
29	0,739	0,001	0,000	0,001	0,860	0,001	0,191	0,001	−0,053	0,805	1,106
30	0,968	0,031	0,001	0,025	0,984	0,013	0,446	0,026	−0,072	0,720	1,168
31	0,104	0,000	0,006	0,000	0,316	0,003	0,258	0,005	−0,125	0,941	1,047
32	0,104	0,000	0,006	0,000	0,315	0,003	0,249	0,005	−0,054	1,256	0,891

Referências

[1] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[2] A.C. Benabdellah, A. Benghabrit, and I. Bouhaddou. A survey of clustering algorithms for an industrial context. In *Procedia Computer Science*, volume 148, pages 291–302, 2019.

[3] James C. Bezdek. Some new indexes of cluster validity. *IEEE Transactions on Systems, Man, and Cybernetics*, 28(3):301–315, 1998.

[4] Ricardo José Gabrielli Barreto Campello et al. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data*, 10(1):5:1–5:51, 2015. Accessed: 2024-03-12.

[5] A.E. Ezugwu, A.M. Ikotun, O.O. Oyelade, L. Abualigah, J.O. Agushaka, C.I. Eke, and A.A. Akinyelu. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, 110:104743, 2022.

[6] Milton Friedman. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11(1):86–92, 1940.

[7] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, 622:178–210, 2023.

[8] Anil K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

[9] Joseph B. Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, 7(1):48–50, 1956.

[10] L. McInnes, J. Healy, and S. Astels. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 2017.

[11] Peter B. Nemenyi. *Distribution-free Multiple Comparisons*. Ph.d. dissertation, Princeton University, 1963.

[12] Frank Nielsen. On a generalization of the jensen–shannon divergence and the jensen–shannon centroid. *Entropy*, 22(2):221, 2021.

[13] Frank Nielsen. On a variational definition for the jensen-shannon symmetrization of distances based on information radius. *Entropy*, 23(4):464, 2021.

[14] S.B. Soheil, N. Müller, C. Plant, and C. Böhm. Clustering of mixed-type data considering concept hierarchies: problem specification and algorithm. *International Journal of Data Science and Analytics*, 10:233–248, 2020.

[15] G. V. Trunk. A problem of dimensionality: A simple example. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(3):306–307, jul 1979.