Análise de Algoritmo de Particionamento usando a Métrica PREDEP

Francisco Neves Tinoco Junior
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Minas Gerais, Brasil
francisconeves@dcc.ufmg.br
Orientado

Flavio Vinicius Diniz de Figueiredo
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Minas Gerais, Brasil
flaviovdf@dcc.ufmg.br
Orientador

Renato Martins Assunção
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Minas Gerais, Brasil
assuncao@dcc.ufmg.br
Colaborador do trabalho

Vinicius Fernandes dos Santos
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Minas Gerais, Brasil
viniciussantos@dcc.ufmg.br
Colaborador do trabalho

***Abstract*—Este trabalho apresenta um algoritmo baseado em Aglomerative Hierarchical Clustering para segmentação de dados e cálculo da métrica PREDEP. O algoritmo mostrou-se eficaz em diversos cenários, ajustando-se às características específicas das distribuições das variáveis. Os resultados evidenciam a importância do critério de parada e da complexidade do modelo em relação ao tamanho do conjunto de dados. A principal contribuição é a demonstração de que um particionamento adaptativo, guiado por uma função de perda, para melhorar a interpretação e precisão da métrica PREDEP.**

I. Introdução

Estabelecer a relação entre duas variáveis é fundamental para uma variedade de tarefas em ciência de dados, desde a seleção de features até a compreensão do comportamento dos objetos analisados. Para essa finalidade, uma gama de métricas está disponível, desde as clássicas, como Pearson e Spearman, até as mais recentes, como o Coeficiente de Informação Máxima (MIC) [1]. Embora amplamente utilizadas, as métricas clássicas frequentemente apresentam limitações significativas. Por exemplo, o coeficiente de Pearson é capaz de detectar apenas relações lineares, enquanto o coeficiente de Spearman está restrito a identificar apenas relações monotônicas. No entanto, a necessidade de determinar e quantificar relações arbitrárias entre variáveis, tanto funcionais quanto não funcionais, persiste e demanda abordagens mais abrangentes.

O PREDEP emerge como uma métrica inovadora para quantificar a capacidade de predição entre variáveis. Ela quantifica a redução da dispersão de uma variável aleatória dado o conhecimento de outra variável, funcionando como uma medida da capacidade preditiva de uma variável com base na outra. Essa métrica apresenta propriedades interessantes, que serão discutidas em detalhes mais adiante, além de introduzir uma abordagem nova em comparação com outras métricas existentes. No entanto, a estimação do PREDEP em dados reais enfrenta desafios, especialmente devido à influência da técnica de partição do espaço utilizada em seu cálculo. Uma técnica de partição mal ajustada pode resultar em interpretações errôneas do comportamento dos dados e na estimativa da métrica.

II. Métrica

A. Caso Discreto

A base da métrica PREDEP é uma extrapolação de uma medida similar, mas aplicada a variáveis discretas. Utilizaremos um exemplo para explicar os conceitos por trás dessa estimativa. Considere uma situação com variáveis aleatórias Y e X, cujas probabilidades são descritas na tabela abaixo.

	Y			
X	Vermelho	Verde	Azul	Total
Urna 1	3/12	0	0	1/4
Urna 2	1/12	1/12	1/12	1/4
Urna 3	0	3/12	0	1/4
Urna 4	0	0	3/12	1/4
Total	1/3	1/3	1/3	

TABLE I
Distribuição conjunta de X e Y

Imagine um cenário onde temos um objeto que segue a distribuição marginal de Y e precisamos prever seu valor. Qual é a probabilidade de acertarmos? Dado que escolhemos uma y_i, a probabilidade de acerto é $P(\text{Acertar Y}|\text{Escolheu } y_i) = P(y_i)$. Portanto, a probabilidade de acerto geral é

$$P(\text{Acertar Y}) = \sum_i P(\text{Acertar Y}|\text{Escolher } y_i) \cdot P(y_i)$$

Em um cenário onde a probabilidade de escolhermos y_i é baseada na probabilidade do evento ocorrer, temos

$$P(\text{Acertar Y}) = \sum_i P(y_i)^2$$

Essa probabilidade é uma medida da dispersão da distribuição dos valores em y. Nesse exemplo, teríamos $P(\text{Acertar X}) = \frac{1}{4}$ e $P(\text{Acertar Y}) = \frac{1}{3}$.

Podemos também condicionar a previsão do valor de Y pelo conhecimento de uma outra variável X. Esse condicionamento pode ser escrito como

$$P(\text{Acertar Y}|X) = \sum_i P_Y(\text{Acertar Y}|X = x_i) \cdot P_X(x_i)$$

O resultado disso é uma métrica que indica a dispersão da distribuição de Y dado que sabemos a priori o valor de X. Na tabela acima, temos que $P(\text{Acertar X}|Y) = 0.625$ e $P(\text{Acertar Y}|X) = 0.83$.

A partir dessas duas métricas, $P(\text{Acertar Y})$ e $P(\text{Acertar Y}|X)$, podemos derivar uma métrica comparativa entre os valores,

$$\tau_{Y|X} = \frac{P(\text{Acertar Y}|X) - P(\text{Acertar Y})}{P(\text{Acertar Y}|X)}$$

$\tau_{Y|X}$ indica o quanto conhecer uma variável X altera a distribuição de Y em comparação à distribuição marginal.

Nesse exemplo, temos:

$$\tau_{X|Y} = \frac{0.625 - 0.25}{0.625} = 0.25 \quad \text{e} \quad \tau_{Y|X} = \frac{0.83 - 0.33}{0.83} = 0.60$$

Esses valores podem ser interpretados da seguinte maneira. No primeiro caso, $X|Y$, conhecer o valor de Y aumenta em 25% a chance de predizer o valor de X. No segundo caso, conhecer o valor de X ajuda a predizer o valor de Y em 60%. Esses valores demonstram uma propriedade interessante dessa métrica: a assimetria. Nesse exemplo, saber o valor de X contribui mais para a previsão de Y do que saber o valor de Y contribui para a previsão de X.

B. Métrica PREDEP

O PREDEP surge como uma generalização dessa medição para variáveis contínuas. Neste trabalho, será utilizada a versão com somente duas variáveis aleatórias, X e Y. Da mesma forma que no caso discreto, o cálculo do PREDEP depende de duas estimativas: uma em relação à marginal (Equação 1) e outra condicionada (Equação 2).

$$S_Y = \mathbb{E}\left[f_Y(Y)\right] = \int f_Y^2(y)\, dy \tag{1}$$

$$\begin{aligned} S_{Y|X} &= \int \left[\int f^2_{Y|X=x}(Y)\, dy \right] f_X(x)\, dx \\ &= \mathbb{E}_X\left[\mathbb{E}_{Y|X}\left[f_{Y|X}(Y|X)\right]\right] . \end{aligned} \tag{2}$$

Diferentemente das suas versões discretas, os valores das duas equações não são diretamente interpretáveis, dependendo, por exemplo, da escala em que estão as variáveis. Isso resulta em uma mudança no intervalo dos valores dessa métrica, que passa de $(0, 1)$ no caso discreto para $(0, \infty)$ no caso contínuo. Portanto, o τ emerge como uma medida capaz de combinar essas duas medições em um valor interpretável.

$$\alpha_{Y|X} = \frac{S_{Y|X} - S_Y}{S_{Y|X}} \tag{3}$$

No trabalho em que o PREDEP é definido, são provadas as seguintes propriedades da métrica:

1) **Não negatividade:** $\alpha_{Y|X} \geq 0$. Isso decorre da noção de que conhecer uma variável não pode aumentar a dispersão da outra, apenas reduzir, o que implica em $S_{Y|X} \geq S_Y$.
2) **Independência:** $\alpha = 0$ se e somente se X e Y são independentes. Essa é uma propriedade extremamente importante da métrica, pois é um argumento a favor de que a métrica captura noções de dependência entre as variáveis.
3) **Intervalo entre 0 e 1:** $\alpha \in [0, 1]$. O valor da métrica indica o quanto conhecer o valor de uma variável reduz a dispersão da outra. Por exemplo, um PREDEP de 0,5 indicaria uma redução de metade da dispersão de Y dado que se conhece X.
4) **Assimetria:** $\alpha_{X|Y} \neq \alpha_{Y|X}$. Essa é uma propriedade bastante interessante do PREDEP, que a diferencia de outras medidas. Essa assimetria traz uma noção de desigualdade entre variáveis, i.e., uma variável pode ter uma relação forte com a outra, contudo o inverso não ser verdadeiro. Por exemplo, considere uma função não invertível $y = x^2$. Nesse caso, x dá toda a informação necessária para se obter y, pois é uma função, contudo y não dá toda a informação para se obter o valor de x, visto que não seria possível saber se o valor original de x era positivo ou negativo.

III. Algoritmo Proposto

O objetivo central deste trabalho é aprimorar a estimativa da métrica, melhorando a estimativa da Equação 2. A estimativa dessa parte em dados é feita por meio de um particionamento dos dados, que deve resultar em uma boa aproximação da condicional $f_{Y|X=x_i}$. A aproximação desse valor por um particionamento do espaço está descrita na Equação 4.

$$S_{Y|X} \approx \sum_i S_{P_{i,X}} \cdot \frac{|P_{i,X}|}{\sum_j |P_{j,X}|} \tag{4}$$

Aonde $P_{i,X}$ representa um particionamento i dos pontos em relação à X e $\frac{|P_{i,X}|}{\sum_j |P_{j,X}|}$ representa a proporção de pontos no particionamento i em relação ao número total de pontos. Na equação apresentada, aproximamos $S_{Y|X}$ como uma soma ponderada dos valores $S_{P_{imX}}$. Aqui, $S_{P_{i,X}}$ representa a estimativa de S_Y aplicada exclusivamente aos pontos contidos na

partição $P_{i,X}$. Portanto, o foco deste trabalho é desenvolver um método eficiente para determinar as partições P_X.

Este desafio de determinar as partições P_X pode ser entendido como um problema de clusterização de dados, onde o objetivo é atribuir rótulos a um conjunto de pontos. No entanto, uma distinção significativa surge em relação aos casos tradicionais de aprendizado não supervisionado: não existe uma noção espacial bem definida nos particionamentos, ou seja, não é necessário que cada partição seja contínua.

A complexidade computacional desse problema decorre da necessidade de examinar todas as combinações possíveis de divisões para encontrar a solução ótima. Essa abordagem resulta em um problema $NP - Hard$, com complexidade assintótica pelo menos fatorial. Portanto, o algoritmo acaba sendo uma heurística para esse problema.

A Figura 2 demonstra por que o particionamento é crucial para esse problema. Com o aumento do número de particionamentos, a métrica tende a aumentar. De fato, à medida que se ganha informação ao aumentar o número de partições, a métrica ou aumenta ou permanece igual. Isso ocorre porque partições mais finas permitem capturar de forma mais detalhada a estrutura dos dados, revelando nuances que podem não ser percebidas em partições mais grosseiras.

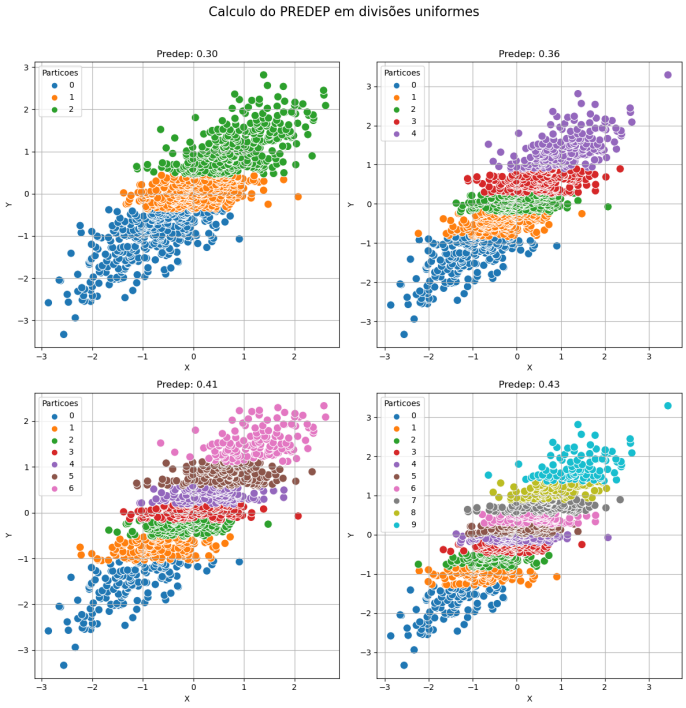

Fig. 1. Diferentes resultados do PREDEP dado o número de partições

Contudo, não se pode aumentar indiscriminadamente o número de partições, pois isso implica uma diminuição da precisão do cálculo da métrica. Além disso, no limite em que se tem um ponto por partição, o valor do PREDEP tende a 1, mesmo em contextos de variáveis independentes. Isso demonstra as complexidades de estimar a métrica em dados reais, pois uma alta granularidade pode levar a uma falsa percepção de dependência entre as variáveis.

Além disso, deve-se considerar em quais regiões os dados serão particionados. A escolha das regiões de particionamento pode influenciar significativamente o valor da métrica PREDEP. Partições que respeitem a distribuição dos dados, capturando adequadamente as áreas de alta e baixa densidade, tendem a fornecer estimativas mais precisas e confiáveis.

Pensando nesses problemas, desenvolvemos um algoritmo baseado em Agglomerative Hierarchical Clustering. Basicamente, o algoritmo substitui a distância por uma função de perda. Além disso, ele só computa a união de partições adjacentes, reduzindo consideravelmente o espaço de busca. A justificativa para essa alteração ser válida é que é possível simular as partições sem essa restrição, i.e., sem a necessidade de um espaço contínuo, utilizando partições com as restrições, mas com um número maior de divisões. Isso resulta em um algoritmo mais rápido, embora com resultados um pouco mais difíceis de interpretar, devido ao maior número de partições.

A figura 2 ilustra o funcionamento do algoritmo. Inicialmente, as partições são geradas com um algoritmo simples de clustering; no nosso caso, utilizamos um particionamento proporcional em relação a X, ou seja, definem-se n partições com o mesmo número de pontos. A cada passo do algoritmo, avalia-se qual união de partições resultará na menor redução da estimativa da métrica, ou seja, a com menor perda. Esse processo continua até que um critério de parada seja atingido.

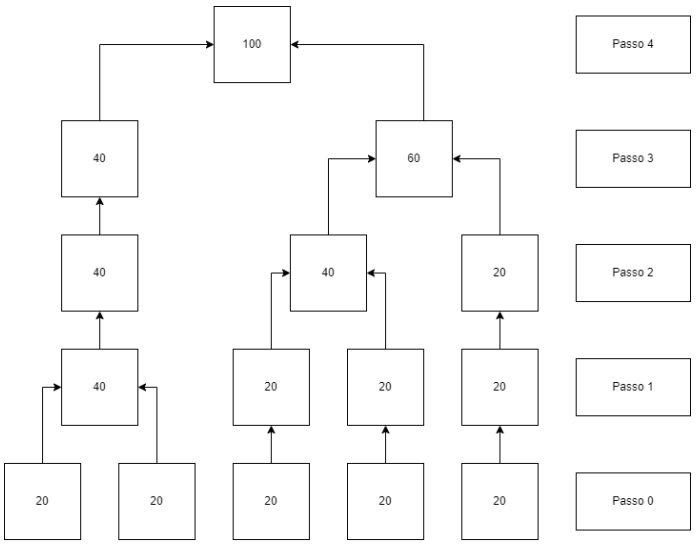

Fig. 2. Diferentes resultados do PREDEP dado o número de partições

Esse algoritmo apresenta algumas vantagens:

1) O espaço resultante se relaciona com a distribuição conjunta das variáveis, não apenas com a distribuição marginal de uma das variáveis. Isso sugere que o particionamento resultante será mais interpretável.
2) É possível controlar e avaliar a complexidade do particionamento resultante, ajustando o nível de granularidade conforme necessário para obter uma estimativa precisa sem perder a interpretabilidade.

3) Analisando quais regiões o algoritmo escolhe unir em cada passo, podemos identificar áreas de interesse nos dados, destacando regiões onde a dependência entre as variáveis é mais pronunciada.
4) O algoritmo tem uma complexidade assintótica linear em relação ao número de partições iniciais sobre a operação do calculo do PREDEP para cada partição, que é a operação mais custosa do algoritmo. Isso é possível devido à restrição de que o espaço seja contínuo.

IV. Metodologia

Para testar o desempenho do algoritmo, utilizamos um framework similar ao de aprendizado de máquina. Nesse framework, temos um conjunto de dados que será utilizado para definir o particionamento e outros conjuntos de dados para calcular a métrica. Para avaliar a robustez da métrica, utilizamos dois conjuntos de teste: um com o mesmo número de pontos que o utilizado para o particionamento e outro com metade desse número. Isso permite avaliar o impacto do número de pontos na estimativa da métrica. O motivo para essa abordagem é obter um resultado mais robusto sobre a métrica, evitando possíveis overfittings na definição das partições. Assim, garantimos que a avaliação do algoritmo seja mais fiel à sua capacidade de generalização, e não apenas ao ajuste aos dados específicos de treinamento.

Para testar diferentes aspectos do comportamento do algoritmo, definimos alguns tipos de relações entre X e Y nas quais conhecemos o resultado teórico de $S_{Y|X}$. A Figura 3 mostra o formato dessas relações. As relações são as seguintes:

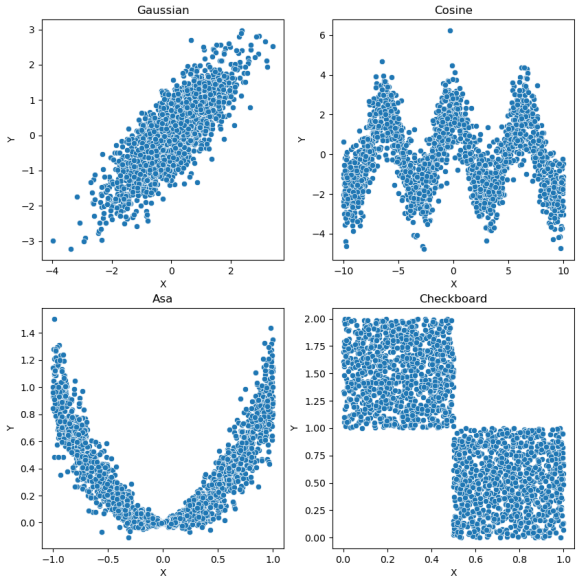

Fig. 3. Formatos das relações entre X e Y utilizadas para testar o algoritmo.

1) **Gaussian:** X e Y seguem uma distribuição gaussiana multivariada com covariância igual a 0.8. Este é um caso simples de uma relação quase linear entre as variáveis, onde esperamos que o algoritmo capture bem a dependência existente.
2) **Cossine:** $X \sim U(0,1)$ e $Y \sim 2\cos(X) + N(0,1)$. Nesta relação, há uma dependência não linear entre X e Y, onde a periodicidade da função cosseno deve ser capturada pelo algoritmo.
3) **Asa:** $X \sim U(-1,1)$ e $Y \sim X^2 + N(0, X/5)$. Este caso é caracterizado por uma variação do erro relacionado ao valor de X, ou seja, há valores de X para os quais o erro é maior e outros para os quais o erro é menor. O algoritmo deve ser capaz de identificar essas áreas de variabilidade distinta.
4) **Checkboard:** $X \sim U(0,1)$ e $Y \sim \begin{cases} U(0,1) & \text{if } X > 0.5 \\ U(1,2) & \text{if } X < 0.5 \end{cases}$. Nesta relação não funcional entre X e Y, a melhor partição é uma divisão pelo valor de 0.5 em X. Este exemplo testa a habilidade do algoritmo de identificar partições ótimas em relações onde não há uma função clara ligando X a Y.

Cada uma dessas relações foi escolhida para testar aspectos específicos do comportamento do algoritmo, desde a captura de dependências lineares e não lineares até a identificação de variações no erro e partições ótimas em cenários não funcionais.

V. Resultados

Nesta seção, apresentamos os resultados da aplicação do algoritmo em diferentes cenários. Dividimos a análise em cada caso específico, mostrando um exemplo do estado julgado mais significativo em cada execução do algoritmo. As figuras associadas ilustram a tendencia do comportamento do algoritmo em cada cenário.

A. Gaussian

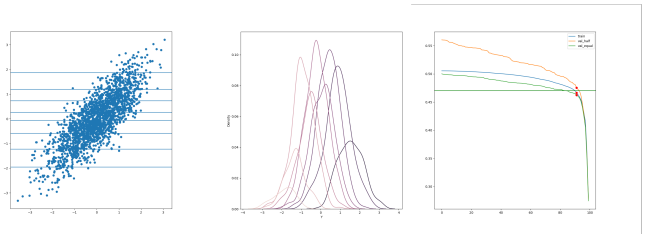

Fig. 4. Particionamento resultante para a relação Gaussian.

Neste caso, o algoritmo converge para uma divisão proporcional do espaço. Ou seja, a cada iteração, ele ajusta as partições para que cada uma tenha a mesma probabilidade. Esta abordagem reflete a propriedade de que todas as variáveis condicionadas possuem o mesmo erro. Portanto, a melhor otimização é criar partições com a mesma probabilidade, garantindo uma distribuição equilibrada entre as partições.

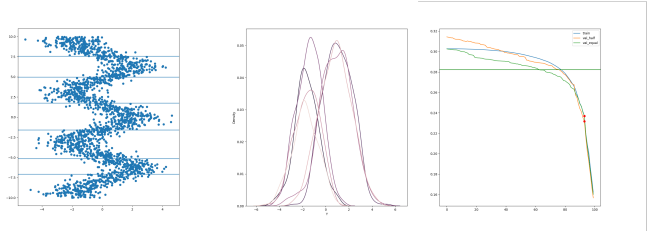

Fig. 5. Particionamento resultante para a relação Cossine.

B. Cossine

O algoritmo converge para divisões nas áreas de pico da função cossenoide, isolando a região central em divisões anteriores. Durante as iterações, o algoritmo ajusta as partições para refletir as regiões com densidades distintas, capturando a periodicidade da relação entre X e Y. O resultado final mostra divisões que correspondem a áreas com distribuições gaussianas distintas, destacando a capacidade do algoritmo de identificar e adaptar-se a padrões complexos na distribuição dos dados.

C. Asa

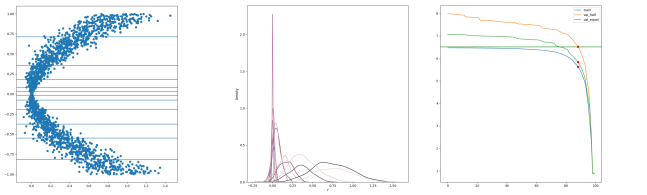

Fig. 6. Particionamento resultante para a relação Asa.

O algoritmo tende a dividir o espaço focando-se nas regiões mais centrais. Neste caso, embora X tenha uma distribuição marginal uniforme, o algoritmo prefere manter mais partições na região central, onde o erro de prever Y dado X é menor. Este comportamento reflete a variação do erro em relação ao valor de X, com o algoritmo identificando e adaptando-se às áreas com diferentes níveis de variabilidade.

D. Checkboard

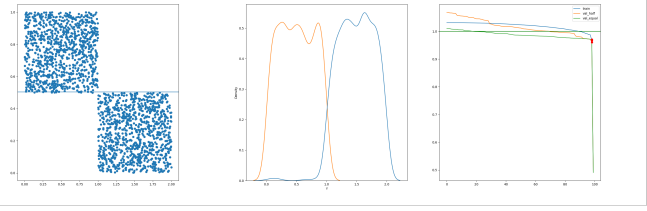

Fig. 7. Particionamento resultante para a relação Checkboard.

O algoritmo converge para a melhor partição, um corte na região central. A figura mostra que o algoritmo identifica corretamente a partição ótima em $X = 0.5$, demonstrando sua eficácia em relações não funcionais onde a dependência entre X e Y é categórica e não contínua.

A partir desses resultados, podemos observar várias características importantes sobre o algoritmo e o problema em questão:

1) **Dependência do critério de parada:** O desempenho do algoritmo em aproximar o valor real da métrica depende consideravelmente do critério de parada. Dada a variedade de relações possíveis na prática, a melhor abordagem seria executar o algoritmo até obter todas as partições e, em seguida, analisar a curva da métrica estimada em relação às partições e aos dados segmentados. Isso ajudaria a determinar o ponto em que a segmentação é satisfatória.
2) **Compreensão do comportamento da distribuição:** Além de calcular o valor da métrica PREDEP, é possível entender o comportamento da distribuição dos dados observando em quais regiões o algoritmo mantém mais partições e quais ele escolhe juntar. Isso fornece insights sobre a estrutura dos dados e suas variabilidades.
3) **Dependência do número de dados:** O valor da estimativa da métrica depende significativamente do número de dados, especialmente quando se utilizam muitas partições. Em todos os exemplos, observa-se que a estimativa com metade do número de pontos resulta em valores mais altos quando há muitas divisões, e os valores das métricas se aproximam quando há menos partições. Assim como no aprendizado de máquina, a complexidade do modelo — que, neste caso, é representada pelo número de partições — deve ser consistente com o tamanho do conjunto de dados. É crucial balancear o número de partições para evitar tanto o overfitting quanto o underfitting, garantindo que a métrica seja calculada de forma precisa e representativa.

VI. Conclusão

Neste trabalho, apresentamos e analisamos um algoritmo baseado em Aglomerative Hierarchical Clustering para a segmentação de dados com o objetivo de calcular a métrica PREDEP. O algoritmo demonstrou eficácia em diversos cenários, adaptando-se às características específicas das distribuições das variáveis envolvidas. Através dos exemplos testados, ficou evidente que o desempenho do algoritmo está intimamente ligado ao critério de parada escolhido, à distribuição dos dados e ao número de partições utilizadas.

Os resultados obtidos destacam a importância de considerar a complexidade do modelo em relação ao tamanho do conjunto de dados, similar ao que é feito em aprendizado de máquina. Além disso, a análise das regiões onde o algoritmo mantém ou junta partições fornece insights valiosos sobre a estrutura dos dados, permitindo uma compreensão mais aprofundada da distribuição conjunta das variáveis.

A principal contribuição deste trabalho é a demonstração de que um particionamento adaptativo, guiado por uma função de perda, que pode melhorar significativamente a interpretação e a precisão da métrica PREDEP. No entanto, é crucial selecionar

um critério de parada adequado para equilibrar a complexidade do particionamento e a representatividade dos dados.

REFERENCES

[1] Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, Lander ES, Mitzenmacher M, Sabeti PC. Detecting novel associations in large data sets. Science. 2011 Dec 16;334(6062):1518-24. doi: 10.1126/science.1205438. PMID: 22174245; PMCID: PMC3325791.