UNIVERSIDADE FEDERAL DE SÃO CARLOS
DEPARTAMENTO DE COMPUTAÇÃO
CURSO DE ENGENHARIA DE COMPUTAÇÃO

YAN GIMENEZ BORGES

Regressão Multi Alvo via Agrupamento Hierárquico das Variáveis Dependentes

São Carlos, SP
07 de agosto de 2024

UNIVERSIDADE FEDERAL DE SÃO CARLOS
DEPARTAMENTO DE COMPUTAÇÃO
CURSO DE ENGENHARIA DE COMPUTAÇÃO

Yan Gimenez Borges

Regressão Multi Alvo via Agrupamento Hierárquico das Variáveis Dependentes

Trabalho de conclusão de curso de graduação apresentado ao Departamento de Computação da Universidade Federal de São Carlos como parte dos requisitos para obtenção do grau de engenheiro da computação.

Orientador: **Prof. Dr. Alexandre Levada**
Co-orientador: **Prof. Dr. Ricardo Cerri**

São Carlos, SP
07 de agosto de 2024

RESUMO

A regressão multi alvo é uma técnica crucial em aprendizado de máquina, aplicada em problemas onde múltiplas variáveis dependentes precisam ser preditas simultaneamente. Neste trabalho, propõe-se uma nova abordagem que utiliza o agrupamento hierárquico das variáveis dependentes para explorar e modelar as relações complexas entre os múltiplos alvos. A metodologia consiste em identificar estruturas hierárquicas subjacentes nos dados de saída, permitindo uma modelagem mais precisa e interpretável das interdependências entre as variáveis dependentes. A abordagem foi avaliada por meio de experimentos extensivos em diversos conjuntos de dados, comparando seu desempenho com métodos tradicionais de regressão multi alvo. A contribuição principal é a introdução de técnicas de agrupamento hierárquico no contexto da regressão multi alvo, proporcionando um arcabouço versátil que pode ser aplicado em diversos domínios. Apesar de bons resultados, análises estatísticas indicam que o uso de agrupamento hierárquico não demonstrou variação estatisticamente significativa na eficácia da regressão multi alvo, o que aponta para a necessidade de utilização de outras métricas de dissimilaridade.

Palavras-chave: regressão. multi alvo. agrupamento hierárquico.

ABSTRACT

Multi-target regression is a crucial technique in machine learning, applied to problems where multiple dependent variables need to be predicted simultaneously. In this work, a new approach is proposed that utilizes hierarchical clustering of the dependent variables to explore and model the complex relationships between the multiple targets. The methodology involves identifying underlying hierarchical structures in the output data, enabling more accurate and interpretable modeling of the interdependencies among the dependent variables. The approach was evaluated through extensive experiments on various datasets, comparing its performance with traditional multi-target regression methods. The main contribution is the introduction of hierarchical clustering techniques in the context of multi-target regression, providing a versatile framework that can be applied across different domains. Despite promising results, statistical analyses indicate that the use of hierarchical clustering did not show statistically significant variation in the efficacy of multi-target regression, highlighting the need for the use of other dissimilarity metrics.

Keywords: regression. multi-target. hierarchical clustering.

LISTA DE ABREVIATURAS E SIGLAS

MTR	Regressão Multi Target
STR	Regressão Single Target
MT	Multi-Target
SOP	Predição de Saída Estruturada
ST	Single Target
DT	Decision Tree
RF	Random Forest
ML	Aprendizado de Máquina
SVR	Support Vector Regressor
RC	Regressor Chain
MTR-HCT	Multi-Target Regression via Hierarchical Clustering of Targets
MTR-RCT	Multi-Target Regression via Random Clustering of Targets

SUMÁRIO

1 INTRODUÇÃO

Regressão é uma área amplamente pesquisada e investigada no aprendizado supervisionado. O objetivo das tarefas clássicas de regressão é aprender, a partir de um conjunto de amostras com saídas alvo conhecidas, uma função que produza um valor contínuo para uma amostra previamente não vista. Na regressão clássica, ou Regressão Single Target (STR), só precisamos mapear as características de entrada para um único alvo de saída, também conhecido como target. No entanto, em Regressão Multi Alvo (MTR) (SPYROMITROS-XIOUFIS et al., 2016), o cenário se torna mais intrincado, pois múltiplos alvos compartilham um conjunto comum de características de entrada e exibem interdependência. Essa interdependência significa que as previsões para um alvo podem influenciar significativamente as previsões de outros alvos.

A regressão tem diversos usos em tarefas reais. Em (KOCEV et al., 2009) é usada para predizer a qualidade e condições da vegetação. Em (MASMOUDI et al., 2020) é utilizado para predizer a concentração de poluentes no ar. Em (HADAVANDI; SHAHRABI; SHAMSHIRBAND, 2015), regressão é aplicado para predizer a produção de energia em fazendas solares e eólicas usando medidas prévias e informações do tempo. Nos últimos anos, regressão tem sido amplamente utilizada em, visão computacional (YAN et al., 2016), análise de imagens médicas (ZHEN et al., 2016) e processamento de linguagem natural (JEONG; LEE, 2009). Além de áreas como ecologia (AHO et al., 2012) e predição de funções gênicas (KOCEV; CECI, 2015).

Comparado com a regressão clássica, o MTR apresenta desafios distintos, resumidos em dois aspectos principais (ZHEN et al., 2018). Primeiramente, há a tarefa de modelar as complexas relações subjacentes entre as características de entrada e múltiplos alvos (MT) de saída. Em segundo lugar, há a necessidade de explorar e compreender a dependência entre os alvos para melhorar a acurácia das predições.

A importância de resolver tarefas de predição de saída estruturada (SOP), especialmente MTR, foi destacada na literatura, reconhecendo seu potencial e importância em diversos domínios. De fato, foi identificada como um dos problemas mais desafiadores em Machine Learning por pesquisadores como (YANG; WU, 2006) e (KRIEGEL et al., 2007). Lidar com as complexidades da MTR envolve navegar pelas intricadas relações entre características de entrada e múltiplos alvos de saída, bem como aproveitar a dependência entre os alvos para aprimorar a precisão das previsões.

Existem duas abordagens principais para o uso de métodos de base no contexto de aprendizado de máquina. A primeira são os métodos de transformação de problema, ou métodos locais, nos quais o problema multi alvo é transformado em vários problemas de STR, cada um resolvido separadamente através de métodos clássicos, como Random Forest (RF) (KOCEV et al., 2009), Boosted Neural Network (HADAVANDI; SHAHRABI;

SHAMSHIRBAND, 2015), Ensembles of tree (KOCEV et al., 2013) e outros. O segundo método é o de algoritmos de adaptação, métodos globais, ou métodos big-bang, que adaptam métodos existentes de single target para prever todas as variáveis-alvo simultaneamente (BORCHANI et al., 2015). Ao utilizar algoritmos de transformação de problema para um domínio de t variáveis-alvo, é necessário construir t modelos preditivos, cada um prevendo uma variável-alvo (KOCEV; CECI, 2015). A previsão para uma amostra não vista seria obtida executando cada um dos t modelos de único destino e concatenando seus resultados. Por outro lado, ao utilizar métodos de adaptação de algoritmo para o mesmo domínio de t variáveis-alvo, apenas um modelo precisaria ser construído, o qual forneceria todas as t previsões.

A literatura demonstra que os métodos de adaptação de algoritmo apresentam um desempenho superior em relação aos métodos de transformação de problema (KOCEV; CECI, 2015) (TSOUMAKAS et al., 2014). A vantagem mais significativa do uso de técnicas MT é que não apenas as relações entre as variáveis da amostra e os alvos são exploradas, mas também as relações entre os alvos entre si. Técnicas single target (ST), por outro lado, eliminam qualquer possibilidade de aprendizado a partir das potenciais relações entre as variáveis de destino, pois um modelo único e independente é treinado para cada alvo separadamente (BEN-DAVID; SCHULLER, 2003).

1.1 OBJETIVO

O objetivo geral desta pesquisa é desenvolver e avaliar abordagens para solucionar o problema de regressão multi-alvo, considerando as correlações entre as variáveis-alvo, de modo a melhorar a precisão das previsões. O trabalho propõe duas metodologias distintas para enfrentar esse desafio, explorando a capacidade de cada uma em capturar as interdependências entre os alvos. As abordagens propostas serão comparadas com modelos de regressão tradicionais amplamente estabelecidos na literatura, buscando evidenciar os ganhos de desempenho decorrentes da incorporação das correlações entre os alvos. Os objetivos específicos desta pesquisa são divididos em duas abordagens:

- Propor uma abordagem de regressão multi alvo via agrupamento hierárquico das variáveis dependentes (MTR-HCT). Isso é feito através da segmentação dos alvos em agrupamentos, utilizando uma matriz de dissimilaridade para fazer um agrupamento hierárquica dos alvos. Ao escolher a melhor partição, por meio do método de silhueta, aplicamos um regressor para cada grupo. Essa proposta, focada diretamente nas relações entre os alvos, visa identificar as correlações das variáveis dependentes durante o processo de agrupamento e aprimorar a performance.

- Propor uma abordagem de regressão multi alvo via agrupamento aleatório das variáveis dependentes (MTR-RCT). O método introduz aleatoriedade na determinação

do número de grupos e na seleção dos alvos que irão compor cada grupo. Em seguida, aplica-se um regressor a cada grupo formado. Posteriormente, os resultados são comparados com os obtidos por meio da abordagem MTR-HCT.

Ambas as abordagens serão avaliadas empiricamente, considerando-se métricas de desempenho comuns em regressão multi-alvo, como média do erro quadrático médio relativo (aRRMSE), Eq. 4.1.

O restante deste trabalho está organizado da seguinte forma: O capítulo 2 descreve brevemente alguns trabalhos relacionados à regressão multi alvo. O capítulo 3 apresenta os detalhes dos métodos propostos MTR-HCT e MTR-RCT para tarefas de regressão multi alvo. Estudos experimentais em uma ampla gama de conjuntos de dados do mundo real são apresentados no capítulo 4. Os resultados e discussões estão agrupados no capítulo 5. Por fim, as conclusões são apresentadas no capítulo 6.

2 FUNDAMENTAÇÃO TEÓRICA

A regressão multi alvo, também conhecida como regressão multivariada, é uma abordagem de aprendizado de máquina que se concentra na predição simultânea de múltiplas variáveis dependentes ou alvos. Ao contrário da regressão univariada, onde o objetivo é prever uma única variável de resposta, a regressão multi alvo busca estimar vários alvos correlacionados ao mesmo tempo. Essa técnica é particularmente útil em contextos onde as variáveis dependentes estão inter-relacionadas e a consideração dessas inter-relações pode melhorar a precisão das previsões.

Segue uma definição mais formal da tarefa de aprendizado de máquina (ML) de regressão multivariada.

Dado:

- Um espaço de entrada X, com tuplas de dimensão d, contendo valores de tipos de dados primitivos, ou seja, $\forall x_i \in X$, $x_i = (x_{i1}, x_{i2}, ..., x_{id})$,
- Um espaço de saída Y (target), com tuplas de dimensão t, contendo valores reais, ou seja, $\forall y_i \in Y$, $y_i = (y_{i1}, y_{i2}, ..., y_{it})$, onde $y_{ik} \in \mathbb{R}$ e $1 \leq k \leq t$,
- Um conjunto de exemplos S, onde cada exemplo é um par de tuplas do espaço de entrada e do espaço de saída, ou seja, $S = (x_i, y_i) | x_i \in X, y_i \in Y, 1 \leq i \leq N$ e N é o número de exemplos em $S (N = |S|)$,
- Um critério de qualidade c, que recompensa modelos com alta precisão.

Encontrar: Uma função $f : X \longrightarrow Y$ tal qual f maximize c. Neste trabalho, a função f funciona como nosso método de regressão utilizado após separar o espaço dos alvos em grupos baseados em suas similaridades.

Existem dois grandes métodos distintos para tratar o problema de regressão multi alvo, o método de transformação de problema (método local) e o método de adaptação de algoritmo (método global). No contexto de transformação do problema para modelos multi alvo, m modelos uni-variado serão treinados no conjunto de dados de forma independentes entre si $D_j = [(x_1^{(1)}, y_j^{(1)}), ..., (x_1^{(N)}, y_j^{(N)})]$, onde $j \in \{1, ..., m\}$.. Essa abordagem, entretanto não permite explorar a relação entre os alvos. Uma outra abordagem para tratar o problema de regressão é o de adaptação de algoritmo, onde será criado apenas um regressor para predizer todos os alvos simultaneamente. Já essa abordagem permite modelar as relações das variáveis dependentes, sendo isso uma grande vantagem dessa abordagem.

A classificação multi-rótulo é um problema de aprendizado supervisionado onde cada instância pode ser associada simultaneamente a múltiplos rótulos ou classes, ao invés

de apenas um único rótulo, como ocorre na classificação tradicional. Formalmente, dado um conjunto de instâncias ($\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$) e um conjunto de rótulos ($\mathbf{L} = \{l_1, l_2, \ldots, l_m\}$), o objetivo da classificação multi-rótulo é aprender uma função $\mathbf{f} : \mathbf{X} \rightarrow 2^{\mathbf{L}}, onde 2^{\mathbf{L}}$ representa o conjunto de todos os subconjuntos possíveis de $\mathbf{L}$, isto é, qualquer instância $x_i \in \mathbf{X}$ pode ser atribuída a um subconjunto de rótulos $S \subseteq \mathbf{L}$. O desafio reside na capacidade de capturar e modelar as possíveis correlações e interdependências entre os rótulos.

Dessa maneira, os métodos de regressão multi-alvo são frequentemente inspirados por técnicas utilizadas na MLC, uma vez que ambas as abordagens compartilham o desafio de prever múltiplos resultados. Assim, algoritmos originalmente desenvolvidos para problemas de MLC, como aqueles baseados em árvores de decisão, redes neurais e regularização de regressão, podem ser adaptados para a regressão multi-alvo, explorando as interdependências entre as saídas para melhorar a precisão das previsões.

2.1 TRABALHOS RELACIONADOS

Regressão multi alvo é uma instância do aprendizado multi tarefa (ZHANG; YANG, 2017), que visa aproveitar informações úteis de múltiplas tarefas relacionadas para melhorar a generalização para todas as tarefas. A principal diferença entre as duas abordagens é que, enquanto o aprendizado multi tarefa pode ter diferentes espaços de entrada, a regressão multi alvo utiliza um espaço de entrada idêntico. Abordagens de multitarefa são aplicáveis para regressões multi alvo, com algumas alterações. A estratégia de regularização convexa aplicada em (ARGYRIOU; EVGENIOU; PONTIL, 2008) (ZHANG; YEUNG, 2012) tem sido amplamente utilizada em tarefas multi alvo. MTR também está intimamente relacionada com Classificação Multi-Rótulo (MLC) (ZHANG; ZHOU, 2014),(PEREIRA et al., 2018). Devido a essa alta correlação, métodos bem estabelecidos de classificação multi-rótulo podem ser quase que diretamente aplicáveis. O stacked single target (SST) (SPYROMITROS-XIOUFIS et al., 2016), ensemble of regressor chains (ERC) (SPYROMITROS-XIOUFIS et al., 2016) e random target combination (RTC) (TSOUMAKAS et al., 2014), sendo este último baseado no método Random K-labelset (Rakel) de MLC (TSOUMAKAS; VLAHAVAS, 2007), são métodos representativos que estendem técnicas bem estabelecidas de classificação multi-label para regressão multi alvo.

Modelar a relação entre entrada e saída e explorar a dependência inter-alvo são os principais desafios da tarefa de regressão multi alvo, os artigos seguintes de MTR trazem diferentes formas para lidar com esses dois desafios. Em (HADAVANDI; SHAHRABI; SHAMSHIRBAND, 2015), invés de utilizar o espaço de entrada completo para treino, foi proposto um novo método de projeção de subespaço baseado em agrupamento (CLSP) para construir automaticamente um espaço de entrada de menor dimensão. Essa abordagem visa tratar espaços de entrada com alta dimensionalidade e melhor modelar a relação

entrada-saída. Utiliza ainda um método de Boosted-Neural Network para lidar com a correlação entre os alvos. De forma semelhante, em (YUAN et al., 2018), propuseram um modelo de seleção de características estruturais esparsas para regressão multi alvo, utilizando uma estrutura de várias camadas com matrizes estruturais e de regressão para reduzir a dimensionalidade e selecionar as variáveis com maior correlação. Já em (MELKI et al., 2017), pensando que a o resultado da predição de uma variável é utilizada como entrada para treinar a próxima variável, foi desenvolvido um método para criar uma matriz de correlação entre todos os alvos e assim utilizar a ordem dos alvos que traria a correlação máxima para construir um regressor chain de Support Vector Regression (SVRCC) e comparando-o com um regressor chain em ordem aleatória (SVRRC). Em (Rahimzadeh Arashloo; KITTLER, 2022), é proposta uma estrutura de aprendizado para capturar as correlações entre os alvos através de um kernel de saída. Ao utilizar uma função de kernel não-linear, é possível explorar as dependências não-lineares entre os alvos.

De forma semelhante a proposta MTR-HCT 3 apresentada nesse trabalho, em (WANG et al., 2019), o autor utiliza um algoritmo de agrupamento hierárquico nas variáveis de saída para identificar as dependências entre os múltiplos alvos, então um método de classificação e de regressão é utilizado para gerar uma matriz de similaridade dependente para cada alvo. Para assim, aprender características específicas dos alvos através das matrizes de similaridade e ao realizar análises de agrupamento das variáveis de entrada.

Outra abordagem, dessa vez de classificação multi-rótulo presente em (GATTO; FERRANDIN; CERRI, 2023), fornece bases interessantes para o desenvolvimento da proposta desse artigo, com métodos em comum como medida de dissimilaridade, agrupamento hierárquico e método de silhueta. Nesse método a autora começa modelando as correlações entre os rótulos usando duas abordagens: a medida de similaridade de Jaccard e uma rede neural auto-organizável de Kohonen (SOM). Com base nessas correlações, o espaço de rótulos é então agrupado em partições híbridas utilizando agrupamento hierárquico aglomerativo. Em seguida, são construídos subconjuntos de dados para cada partição híbrida, com base nas instâncias de treinamento associadas aos rótulos de cada cluster. A validação é realizada por duas estratégias: avaliando o desempenho dos classificadores para cada partição e calculando o coeficiente de silhueta. Por fim, a melhor partição é escolhida com base nos melhores resultados de Macro-F1 e Micro-F1 ou no coeficiente de silhueta, conforme a estratégia de validação utilizada.

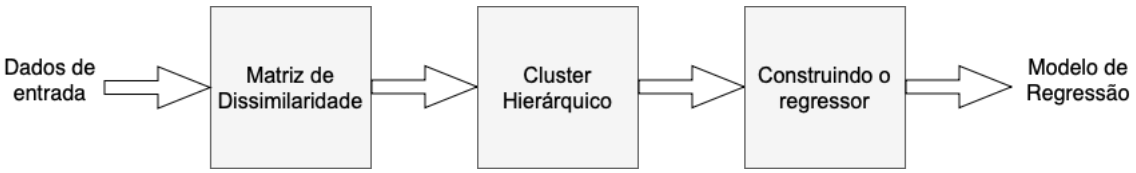
Pelos trabalhos revisados, pode-se observar que a correlação entre as variáveis dependentes são exploradas de diversas formas diferentes. Nessa proposta, a correlação entre os alvos é utilizada para agrupar alvos altamente correlacionados em um mesmo grupo hierárquico, enquanto busca pela melhor partição das variáveis dependentes.

3 PROPOSTA

3.1 MÉTODO MTR-HCT

Nesta seção, apresentamos o método proposto, denominado MTR-HCT, que captura a correlação entre os alvos na tarefa de regressão multi alvo. A arquitetura do método proposto é apresentada na Figura 1. Primeiramente, construímos uma **Matriz de Dissimilaridade** ao calcular a correlação de Pearson para cada par de alvos. Em seguida, aplicamos um algoritmo de agrupamento hierárquico baseado na medida de dissimilaridade entre as variáveis dependentes e determinamos a melhor partição utilizando o método da silhueta. Dessa forma, agrupamos em um mesmo grupo alvos com maior similaridade. Finalmente, um modelo de regressão é implementado para cada grupo, realizando a predição dos alvos de maneira global dentro de cada grupo. A ideia é que o regressor seja implementado levando em conta apenas os dados com alta similaridade entre si. Por isso, o método MTR-HCT é híbrido, posicionando-se entre um método local e um método global: ele não constrói um regressor individual para cada alvo, como faria um método local, nem um único regressor para todos os alvos, tal qual um método global.

Figura 1 – Arquitetura do modelo MTR-HCT



Fonte: Autor

A implementação está disponível publicamente no Github.

3.1.1 Matriz de Dissimilaridade

A matriz de dissimilaridade é construída com base na correlação de Pearson (r) entre duas variáveis X e Y, a qual é dada pela fórmula:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \tag{3.1}$$

onde:

X_i valor individual da variável X
Y_i valor individual da variável Y
$\overline{X}$ média da variável X

$\overline{Y}$ média da variável Y
n número de pares de observações

A dissimilaridade (d) baseada na correlação de Pearson pode ser calculada como:

$$d = 1 - r \tag{3.2}$$

onde:

d dissimilaridade
r correlação de Pearson

Para construir a matriz de dissimilaridade, calculamos a dissimilaridade entre todos os pares de variáveis. Suponha que temos um conjunto de dados com p variáveis. A matriz de dissimilaridade D será uma matriz $p \times p$ onde cada entrada d_{ij} representa a dissimilaridade entre as variáveis i e j.

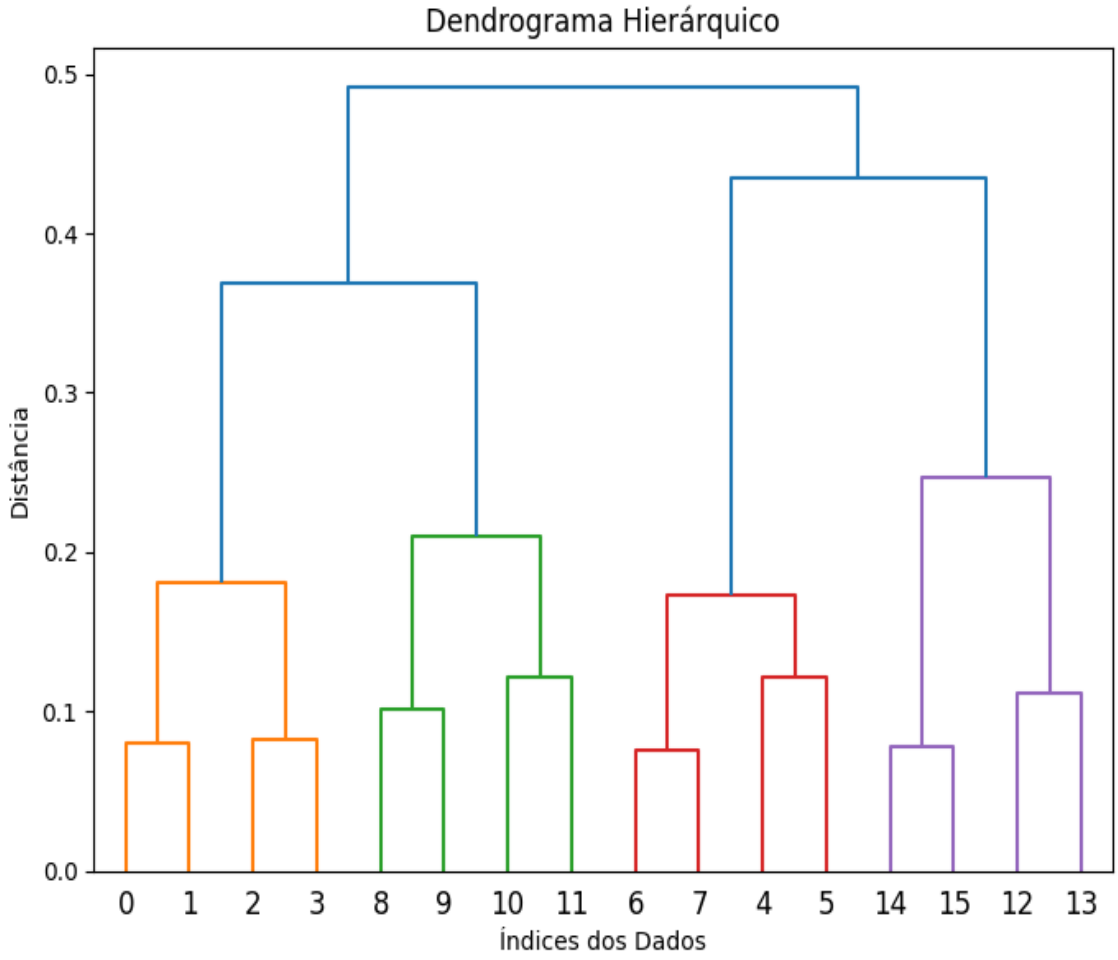
$$D = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1p} \\ d_{21} & d_{22} & \cdots & d_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ d_{p1} & d_{p2} & \cdots & d_{pp} \end{pmatrix}$$

Cada elemento d_{ij} na matriz D é calculado usando as fórmulas mencionadas acima, onde $d_{ij} = 1 - r_{ij}$ e r_{ij} é a correlação de Pearson entre as variáveis i e j. A diagonal da matriz D será zero ($d_{ii} = 0$), pois a dissimilaridade de uma variável consigo mesma é zero.

3.1.2 Agrupamento Hierárquico

Em regressão multi alvo, uma amostra é representada por um vetor de entrada e um vetor de saída de múltiplos alvos. Assumimos que targets mutuamente dependentes dividam características semelhantes no espaço de saída. Neste trabalho, modelamos as similaridades entre os targets através do agrupamento hierárquico sobre o espaço de saída. A Figura 2 ilustra o dendrograma gerado a partir da matriz de dissimilaridade para visualizar o agrupamento dos dados a partir do dataset SCM20D utilizado para análise, presente na Tab. 1.

Figura 2 – Dendrograma do dataset scm20d



Fonte: Autor

A partir da construção do dendrograma, utilizamos o **Método de Silhueta**, uma técnica utilizada para avaliar a qualidade de um agrupamento, ou seja, para medir quão bem os objetos foram agrupados dentro dos grupos. A análise de silhueta fornece uma medida de quão semelhante um objeto é ao seu próprio grupo (coesão) em comparação com outros agrupamentos (separação). Essa medida é utilizada para determinar o número adequado de grupos.

Para cada ponto i em um conjunto de dados, a silhueta é calculada como segue:

1. **Coeficiente de Coesão ($a(i)$)**:
   - $a(i)$ é a média das distâncias entre o ponto i e todos os outros pontos no mesmo grupo. Representa a coesão dentro do grupo.
2. **Coeficiente de Separação ($b(i)$)**:

- $b(i)$ é a menor média das distâncias entre o ponto i e todos os pontos de qualquer outro grupo, do qual i não faz parte. Representa a separação entre grupos.

3. **Coeficiente de Silhueta ($s(i)$)**:

   - O coeficiente de silhueta para o ponto i é dado por:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{3.3}$$

   - O valor de $s(i)$ varia de -1 a 1:
     - $s(i)$ próximo de 1 indica que o ponto está bem agrupado.
     - $s(i)$ próximo de 0 indica que o ponto está na fronteira entre dois grupos.
     - $s(i)$ negativo indica que o ponto pode ter sido mal agrupado.

Dessa forma, é possível selecionar a partição que proporciona a maior correlação entre os targets. Termina-se, então, com c grupos, onde $c \ll t$, sendo t o número de targets. Consequentemente, teremos que construir c regressores globais, cada um utilizando apenas a informação dos targets semelhantes dentro de seu respectivo grupo.

Como o interesse está na qualidade da separação e não em melhorar um regressor específico, utilizamos diversos regressores base, os quais implementamos no nosso modelo para atestar sua eficácia.

3.2 MÉTODO MTR-RCT

Nesta seção, apresentamos o método proposto, denominado MTR-RCT, que realiza uma partição aleatória dos alvos. A arquitetura do método, descrita no Algoritmo 1, envolve definir aleatoriamente o número de agrupamentos, que pode variar de 2 até o número total de alvos menos um. Inicialmente, um alvo é atribuído aleatoriamente a cada cluster, e os alvos restantes são distribuídos aleatoriamente entre os grupos. Esse processo garante que cada agrupamento contenha pelo menos um alvo e que todos os alvos sejam distribuídos entre os grupos. Para assegurar uma média mais precisa dos resultados, o método é iterado 10 vezes, aumentando a variedade de combinações, já que a partição dos dados pode variar significativamente entre as iterações. Finalmente um modelo de regressão é implementado para cada cluster, realizando a predição dos targets através de uma abordagem global dentro de cada grupo. Portanto, trata-se de um método híbrido baseado na aleatoriedade para realizar as partições.

Algorithm 1: Algoritmo de Agrupamento Randômico

Input: Matriz de Targets D

Output: $grupos\_rand$ (lista de grupos aleatório)

$ITER \leftarrow 10$
$n\_target \leftarrow$ número de targets
$grupos\_rand \leftarrow []$
for $i = 1$ ***to*** $ITER$ **do**
    $n\_grupos \leftarrow$ RANDOM_INT$(2, n_target)$;
    $grupos\_target \leftarrow$ Cria uma lista vazia n_grupos vezes;
    $targets \leftarrow [0, 1, \ldots, n\_target - 1]$;
    $targets \leftarrow$ EMBARALHA$(targets)$;

    # Passo 1: Distribui pelo menos um target para cada grupo;
    for $grupo = 0$ ***to*** n_grupos **do**
        $target \leftarrow$ Salva o último elemento de $targets$ e retira da lista;
        $grupo\_target_{grupo} \leftarrow grupo\_target_{grupo} \bigcup D_{target}$;

    # Passo 2: Distribui os targets restantes de forma aleatória;
    while $targets \neq []$ **do**
        $grupo \leftarrow$ RANDOM_INT$(0, n_grupo)$;
        $target \leftarrow$ Salva o último elemento de $targets$ e retira da lista;
        $grupo\_target_{grupo} \leftarrow grupo\_target_{grupo} \bigcup D_{target}$;
    $grupos\_rand_i \leftarrow grupos\_rand_i \bigcup grupos\_target$;

Feito o agrupamento randômico, podemos aplicar os algoritmos de regressão. Contruímos c grupos de forma aleatória, onde $c \ll t$, sendo t o número de targets. Dessa forma, cada regressor prediz de forma global os targets intra cluster.

4 METODOLOGIA

Nesse capítulo, é apresentado o experimento para verificar a efetividade do método proposto em lidar com tarefas de regressão multi-target. Todos os experimentos foram conduzidos em um computador equipado com um chip M1 com 8 núcleos heterogêneos (4 de alta performance e 4 de alta eficiência) e 8GB de RAM.

4.1 MÉTRICAS DE AVALIAÇÃO

A performance preditiva de um modelo MTR é avaliada usando a média do erro quadrático médio relativo (aRRMSE), que calcula a média dos erros médios quadráticos relativos (RRMSE) para as variáveis-alvo individuais. RRMSE é uma medida relativa calculada em relação ao modelo de referência que prevê a média aritmética de todos os valores de uma variável-alvo específica no conjunto de aprendizado. Especificamente, o valor y_i é a previsão real de referência para a i-ésima variável-alvo, enquanto o valor $\hat{y}_i^{(e)}$ representa o valor previsto para a i-ésima variável-alvo do exemplo e.

$$aRRMSE = \frac{1}{t}\sum_{i=1}^{t} RRMSE = \frac{1}{t}\sum_{i=1}^{t}\sqrt{\frac{\sum_{e=1}^{N}(y_i^{(e)} - \hat{y}_i^{(e)})^2}{\sum_{e=1}^{N}(y_i^{(e)} - \bar{y}_i^{(e)})^2}} \tag{4.1}$$

onde $\bar{y}_i^{(e)}$ é a média da variável-alvo y_i, t o número de targets e N o número de exemplos.

A métrica de aRRMSE foi calculada aplicando o método de cross validation com 5-fold para todos os datasets. Com exceção do SCM1D, SCM20D, RF1 e RF2 que devido ao tamanho e limitações computacionais, utilizou-se 2-fold cross validation.

4.2 CONJUNTO DE DADOS E SETUP DO EXPERIMENTO

Para verificar a eficácia do método proposto, foram adotados 16 exemplos reais de bases de dados multi alvo retirados do Mulan [1], que tem como fonte (SPYROMITROS-XIOUFIS et al., 2016). A escolha desses A Tabela 1 mostra os datasets utilizados em nosso experimento, junto com suas estatísticas, onde N é o número de amostras, d é o número de variáveis de entrada e a é o número de alvos. Os 16 conjuntos de dados estão em ordem alfabética. As bases com dados faltantes foram substituídos com a média da amostra de dados para aquela variável.

O método proposto MTR-HCT foi comparado contra várias abordagens de regressão multi alvo de ponta para validar sua performance de predição, bem como comparar com o método de agrupamento randômico MTR-RCT para verificar a eficácia da partição dos alvos.

[1] http://mulan.sourceforge.net/datasets-mtr.html

Tabela 1 – Estatística dos 16 datasets usados no experimento

Dataset	Nome Completo	Amostras (N)	Entrada (d)	Alvos (a)
ANDRO	Andromeda	49	30	6
ATP1D	Airline Ticket Price 1	337	411	6
ATP7D	Airline Ticket Price 2	296	411	6
EDM	Electrical Discharge Machine	154	16	2
ENB	Energy Building	768	8	2
JURA	Jura	359	15	3
OES10	Occupational Employment Survey 1	403	298	16
OES97	Occupational Employment Survey 2	334	263	16
OSALES	Online Sales	639	413	12
RF1	River Flow 1	9125	64	8
RF2	River Flow 2	9125	576	8
SCM1D	Supply Chain Management 1	9803	280	16
SCM20D	Supply Chain Management 2	8966	61	16
SCPF	See Click Predict Fix	1137	23	3
SLUMP	Concrete Slump	103	7	3
WQ	Water Quality	1060	16	14

Tabela 2 – Regressores implementados

Abreviação	Nome Completo
DT	Decision Tree
RF	Random Forest
RC-DT	Regressor Chain de Decision Tree
RC-RF	Regressor Chain de Random Forest
XGB-G	XGBoost Global
XGB-L	XGBoost Local

Para cada regressor, foi implementada uma versão original dele, uma versão em nosso método MTR-HCT e outra pelo método MTR-RCT.

5 EXPERIMENTOS E DISCUSSÃO

5.1 RESULTADOS

A seguir, na Tabela 3 segue o resultado do agrupamento em cada dataset, indicando o número de alvos e como foi feito o agrupamentos dos alvos.

Tabela 3 – Agrupamento dos 16 datasets usados no experimento

Dataset	N.º alvos	Agrupamento
ANDRO	6	(2,2,2)
ATP1D	6	(3,3)
ATP7D	6	(2,2,1,1)
EDM	2	(2)
ENB	2	(2)
JURA	3	(3)
OES10	16	(15,1)
OES97	16	(15,1)
OSALES	12	(11,1)
RF1	8	(2,2,1,1,1,1)
RF2	8	(2,2,1,1,1,1)
SCM1D	16	(2,2,2,2,2,2,2,2)
SCM20D	16	(2,2,2,2,2,2,2,2)
SCPF	3	(3)
SLUMP	3	(3)
WQ	14	(5,9)

Os resultados das predições em função de aRRMSE do nosso método proposto HCT e RCT e dos métodos clássicos, estão nas Tabelas 4 e 5, onde os melhores resultados de cada dataset estão em negrito. Quanto menor o valor melhor.

Tabela 4 – Resultado de 5 datasets usados no experimento

Regressor	ANDRO	SCM1D	SCM20D	WQ	OSALES
Partição	(2,2,2)	(2,2,2,2,2,2,2,2)	(2,2,2,2,2,2,2,2)	(5,9)	(11,1)
DT	0,9069	0,7120	0,7907	1,1252	1,0162
HCT DT	0,7613	0,6851	0,7786	1,1223	1,0120
RCT DT	0,7912	0,6950	0,7849	1,1293	1,0144
RF	0,7416	0,5772	**0,6381**	0,9489	**0,8540**
HCT RF	0,7450	**0,5697**	0,6419	0,9509	0,8544
RCT RF	0,7493	0,5727	0,6403	0,9541	0,8630
RC-DT	0,7503	0,6998	0,8391	1,1260	1,0476
HCT RC-DT	0,7683	0,6924	0,7957	1,1181	1,0373
RCT RC-DT	0,8413	0,6862	0,7926	1,1252	1,0131
RC-RF	0,7577	0,5821	0,6732	0,9568	0,8878
HCT RC-RF	0,7481	0,5751	0,6497	0,9561	0,8690
RCT RC-RF	0,7624	0,5706	0,6466	0,9553	0,8651
XGB-L	0,6922	0,5891	0,7214	0,9509	0,9028
XGB-G	0,7112	0,6274	0,7520	**0,9465**	0,9103
HCT XGB-G	**0,6789**	0,5928	0,7243	0,9469	0,9238
RCT XGB-G	0,6858	0,6002	0,7292	0,9506	0,9002

Tabela 5 – Resultado de 6 datasets usados no experimento

Regressor	ATP1D	ATP7D	RF1	RF2	OES10	OES97
Partição	(3,3)	(2,2,1,1)	(2,2,1,1,1,1)	(2,2,1,1,1,1)	(15,1)	(15,1)
DT	0,7278	0,8404	0,3517	0,3333	0,7798	0,8786
HCT DT	0,7037	0,8438	**0,2573**	0,3237	0,8272	0,8290
RCT DT	0,7180	0,8341	0,2809	0,3436	0,7697	0,8204
RF	0,6284	0,7526	0,3542	0,4435	0,6462	0,7150
HCT RF	**0,6038**	0,7227	0,2938	0,3222	0,6787	0,7519
RCT RF	0,6112	0,7390	0,2638	0,3434	0,6224	0,6942
RC-DT	0,7243	0,8699	0,2701	0,3476	0,7602	0,8170
HCT RC-DT	0,7180	0,8263	0,2622	0,3208	0,8404	0,8540
RCT RC-DT	0,7130	0,8560	0,2807	**0,3098**	0,7635	0,8093
RC-RF	0,6168	0,7646	0,3230	0,3937	0,6230	0,6909
HCT RC-RF	0,6122	0,7258	0,2982	0,3256	0,6640	0,7364
RCT RC-RF	0,6102	0,7439	0,2605	0,3121	**0,6168**	**0,6862**
XGB-L	0,6100	0,7571	0,4174	0,5496	0,6477	0,7341
XGB-G	0,6041	0,6916	0,4975	0,5401	0,6619	0,7737
HCT XGB-G	0,6115	0,7114	0,4353	0,4860	0,7093	0,7691
RCT XGB-G	0,6086	0,7279	0,3996	0,4547	0,6454	0,7318

Como o modelo é de agrupamento dos alvos, para bases de dados com poucos variáveis dependentes, não é possível rodar o algoritmo, visto que o método de agrupamento hierárquico automaticamente identifica que é melhor não separar em múltiplos clusters. Dessa forma, não há resultados para as bases Edm, Enb, Jura, Scpf e Slump.

A partir das Tabelas 4 e 5, podemos observar que o método HCT apresentou o melhor resultado em 4 datasets, enquanto o método RCT obteve o melhor desempenho em 3 datasets. No entanto, ficou evidente uma limitação do método HCT em casos como dos datasets OES10 e OES97. Nessas situações, o agrupamento hierárquico separa os clusters de forma muito concentrada, onde muitos alvos são agrupados em um único cluster, enquanto apenas um alvo é colocado em outro cluster. Esse tipo de segmentação provou ser bastante prejudicial para a performance do método HCT, resultando em um desempenho inferior em comparação com os outros dois métodos.

Em contrapartida, nos datasets onde o método HCT teve um desempenho superior, o agrupamento hierárquica segmentou os dados em clusters de tamanhos semelhantes, favorecendo assim a eficácia do método.

5.2 ANÁLISE ESTATÍSTICA

Foi aplicado o teste de Friedman para análise comparativa dos algoritmos. A capacidade do teste de detectar diferenças estatisticamente significativas entre os algoritmos, proporciona uma análise estatística robusta e confiável para determinar se há diferenças significativas no desempenho dos algoritmos avaliados.

Tabela 6 – Teste de Friedman

Estatística de Friedman	0,8323
Valor crítico ($\alpha = 0,05$)	0.99
Aceito ou Rejeitado H_0	Aceito

Com o valor-p de 0.99, maior que o nível de significância de 0.05, não há evidências suficientes para rejeitar a hipótese nula. Portanto não podemos considerar que há diferenças significativas no desempenho dos algoritmos testados.

5.3 DISCUSSÃO

No estudo presente, a aplicação do teste estatístico de Friedman revelou que não há evidências suficientes para rejeitar a hipótese nula de que não existem diferenças significativas entre os métodos de agrupamento hierárquico avaliados. Este resultado sugere que, dentro do escopo deste estudo e das condições experimentais estabelecidas, os métodos comparados não demonstraram variação estatisticamente significativa em sua eficácia. Essa constatação levanta a possibilidade de que a formulação do algoritmo de agrupamento hierárquico, tal como aplicada neste estudo, pode não ser sensível o suficiente para

capturar as correlações entre os targets. Cabe em trabalhos futuros experimentar com o algoritmo responsável por identificar a correlação entre os targets, como por exemplo, o uso de uma função de kernel de saída não-linear pode ser uma abordagem que traga resultados interessantes.

6 CONCLUSÃO

Neste trabalho, apresentamos uma abordagem inovadora para a regressão multi-alvo, utilizando o agrupamento hierárquico das variáveis dependentes como meio de explorar e modelar as relações complexas entre múltiplos alvos simultaneamente. A metodologia baseia-se na premissa de que a identificação de estruturas subjacentes nos dados de saída pode levar a modelos mais precisos e interpretáveis, aproveitando ao máximo a informação contida nas interdependências dos alvos.

Os experimentos realizados com diversos conjuntos de dados demonstraram que o uso de agrupamento hierárquico das variáveis dependentes, apesar de ser promissor, requer maiores estudos acerca de como mediar a dependência entre as variáveis dependentes. A utilização do coeficiente de correlação de Pearson, que é uma medida linear de dependência entre duas variáveis aleatórias, e por essa razão, é menos sensível a correlações não-lineares das variáveis-alvo e outliers, figura entre possibilidades de trabalhos futuros.

Há ainda várias outras direções para trabalhos futuros. Primeiramente, a escolha do método de agrupamento pode ter um impacto significativo nos resultados, e uma investigação mais aprofundada sobre como otimizar essas escolhas para diferentes tipos de dados seria benéfica. Outro aspecto importante a ser explorado é a aplicação da metodologia proposta em problemas com um número muito grande de alvos, onde a escalabilidade do método pode se tornar um desafio. Estratégias de simplificação e aproximação, bem como a utilização de algoritmos paralelos e distribuídos, são caminhos promissores para superar essas limitações.

Em conclusão, a regressão multi-alvo via agrupamento hierárquico das variáveis dependentes oferece uma abordagem versátil para lidar com problemas complexos de predição com múltiplos alvos. Sendo assim, este trabalho abre possibilidades para pesquisas futuras e aplicações práticas, contribuindo para o desenvolvimento contínuo de métodos mais eficientes e interpretáveis em aprendizado de máquina e análise de dados.

APÊNDICES

APÊNDICE A – DATASETS

Andro

O conjunto de dados Andromeda (HATZIKOS et al., 2008) trata da previsão de valores futuros para seis variáveis de qualidade da água (temperatura, pH, condutividade, salinidade, oxigênio, turbidez) no Golfo de Thermaikos, em Thessaloniki, Grécia. As medições das variáveis-alvo são feitas por sensores subaquáticos com um intervalo de amostragem de 9 segundos e, em seguida, são calculadas as médias para obter uma única medição de cada variável ao longo de cada dia. O conjunto de dados específico que usamos aqui corresponde ao uso de uma janela de 5 dias (ou seja, os atributos das características correspondem aos valores das seis variáveis de qualidade da água até 5 dias no passado) e um intervalo de 5 dias (ou seja, prevemos os valores de cada variável 6 dias à frente).

Atp

O conjunto de dados Airline Ticket Price (SPYROMITROS-XIOUFIS et al., 2016) diz respeito à previsão de preços de passagens aéreas. As linhas são uma sequência de observações ordenadas no tempo ao longo de vários dias. Cada amostra neste conjunto de dados representa um conjunto de observações de um par específico de data de observação e data de partida. As variáveis de entrada para cada amostra são valores que podem ser úteis para a previsão dos preços das passagens aéreas para uma data de partida específica. As variáveis alvo nesses conjuntos de dados são o preço do dia seguinte (ATP1D) ou o preço mínimo observado nos próximos 7 dias (ATP7D) para 6 preferências de voo alvo: (1) qualquer companhia aérea com qualquer número de escalas, (2) qualquer companhia aérea sem escalas, (3) Delta Airlines, (4) Continental Airlines, (5) Airtran Airlines e (6) United Airlines.

As variáveis de entrada incluem os seguintes tipos: o número de dias entre a data de observação e a data de partida (1 característica), variáveis booleanas para o dia da semana da data de observação (7 características), a enumeração completa dos seguintes 4 valores: (1) o preço mínimo, preço médio e número de cotações de (2) todas as companhias aéreas e de cada companhia aérea que cotou mais de 50% dos dias de observação (3) para voos sem escalas, com uma escala e com duas escalas, (4) para o dia atual, dia anterior e dois dias anteriores. O resultado é um conjunto de características com 411 variáveis. Para detalhes específicos sobre como esses conjuntos de dados são construídos, consulte (GROVES; GINI, 2015). A natureza desses conjuntos de dados é heterogênea, com uma mistura de vários tipos de variáveis, incluindo variáveis booleanas, preços e contagens.

Edm

O conjunto de dados Electrical Discharge Machining (KARALIČ; BRATKO, 1997)

representa um problema de regressão com dois alvos. A tarefa é reduzir o tempo de usinagem reproduzindo o comportamento de um operador humano que controla os valores de duas variáveis. Cada uma das variáveis-alvo assume 3 valores numéricos distintos (-1, 0, 1) e há 16 variáveis de entrada contínuas.

Enb

O conjunto de dados Energy Building (TSANAS; XIFARA, 2012) diz respeito à previsão das necessidades de carga de aquecimento e carga de resfriamento de edifícios (ou seja, eficiência energética) em função de oito parâmetros do edifício, como área de envidraçamento, área do telhado e altura total, entre outros.

Jura

O conjunto de dados Jura (GOOVAERTS, 1997) consiste em medições das concentrações de sete metais pesados (cádmio, cobalto, cromo, cobre, níquel, chumbo e zinco), registradas em 359 locais na camada superficial do solo de uma região do Jura Suíço. O tipo de uso do solo (Floresta, Pastagem, Prado, Cultivo) e o tipo de rocha (Argoviano, Kimmeridgiano, Sequaniano, Portlandiano, Quaternário) também foram registrados para cada local. Em um cenário típico (GOOVAERTS, 1997), estamos interessados na previsão da concentração de metais que são mais caros de medir (variáveis primárias) usando medições de metais que são mais baratos de amostrar (variáveis secundárias). Neste estudo, cádmio, cobre e chumbo são tratados como variáveis alvo, enquanto os metais restantes, juntamente com o tipo de uso do solo, tipo de rocha e as coordenadas de cada local são usados como características preditivas.

Oes

Os conjuntos de dados Occupational Employment Survey foram obtidos dos anos 1997 (OES97) e 2010 (OES10) da Pesquisa de Emprego Ocupacional anual compilada pelo Bureau of Labor Statistics dos EUA. Cada linha fornece o número estimado de empregados equivalentes a tempo integral em vários tipos de emprego para uma área metropolitana específica. Há 334 e 403 cidades nos conjuntos de dados de 1997 e 2010, respectivamente. As variáveis de entrada nesses conjuntos de dados são um subconjunto sequenciado aleatoriamente dos tipos de emprego (por exemplo, médico, dentista, técnico de reparação de automóveis, etc.) observados em pelo menos 50% das cidades (algumas categorias não tinham valores para determinadas cidades). Os alvos para ambos os anos são selecionados aleatoriamente do conjunto total de categorias acima do limite de 50%. Valores ausentes tanto nas variáveis de entrada quanto nas variáveis-alvo foram substituídos pelas médias amostrais para esses resultados.

Osales

Esta é uma versão pré-processada do conjunto de dados usado na competição "Online Product Sales"do Kaggle (CHUDZICKI, 2012), que diz respeito à previsão de vendas online de produtos de consumo. Cada linha no conjunto de dados corresponde a um produto diferente, descrito por várias características do produto, bem como por características de uma campanha publicitária. Existem 12 variáveis alvo correspondentes às vendas mensais dos primeiros 12 meses após o lançamento do produto.

Rf

Os conjuntos de dados River Flow (SPYROMITROS-XIOUFIS et al., 2016) dizem respeito à previsão dos fluxos da rede fluvial para 48 horas no futuro em locais específicos. O conjunto de dados contém dados de observações horárias de fluxo para 8 locais na rede do Rio Mississippi, nos Estados Unidos, obtidos do Serviço Nacional de Meteorologia dos EUA. Cada linha inclui a observação mais recente para cada um dos 8 locais, bem como observações defasadas no tempo de 6, 12, 18, 24, 36, 48 e 60 horas no passado. No RF1, cada local contribui com 8 variáveis de atributo para facilitar a previsão. Há um total de 64 variáveis, além de 8 variáveis alvo. O conjunto de dados RF2 estende os dados do RF1 adicionando informações de previsão de precipitação para cada um dos 8 locais (chuva esperada relatada como valores discretos: 0.0, 0.01, 0.25, 1.0 polegadas). Para cada observação e local de medição, a previsão de precipitação para janelas de 6 horas até 48 horas no futuro é adicionada (6, 12, 18, 24, 30, 36, 42 e 48 horas). Ambos os conjuntos de dados contêm mais de um ano de observações horárias (>9000 horas) coletadas de setembro de 2011 a setembro de 2012. O domínio é um candidato natural para regressão multi-alvo porque há claras relações físicas entre as leituras na rede fluvial contígua.

SCM

Os conjuntos de dados de Gerenciamento da Cadeia de Suprimentos são derivados do torneio Trading Agent Competition in Supply Chain Management (TAC SCM) de 2010. Os métodos precisos para pré-processamento e normalização dos dados são descritos em detalhes por (GROVES; GINI, 2015). Alguns valores de referência para a precisão da previsão neste domínio estão disponíveis no TAC SCM Prediction Challenge de 2008, e esses conjuntos de dados correspondem apenas ao tipo de previsão "Product Future". Cada linha corresponde a um dia de observação no torneio (há 220 dias em cada jogo e 18 jogos no torneio). As variáveis de entrada nesse domínio são os preços observados para um dia específico do torneio. Além disso, 4 observações com atraso no tempo são incluídas para cada produto e componente observado (atrasos de 1, 2, 4 e 8 dias) para facilitar a antecipação de tendências futuras. Os conjuntos de dados contêm 16 alvos de regressão, cada alvo corresponde ao preço médio do dia seguinte (SCM1D) ou ao preço médio para 20 dias no futuro (SCM20D) para cada produto na simulação. Dias sem valores-alvo são

excluídos dos conjuntos de dados (ou seja, dias com rótulos que estão além do fim do jogo são excluídos).

SCPF

Esta é uma versão pré-processada do conjunto de dados usado na competição "See Click Predict Fix" do Kaggle (CUKIERSKI, 2013). Diz respeito à previsão de três variáveis alvo que representam o número de visualizações, cliques e comentários que um problema específico do 311 receberá. Os problemas foram coletados de 4 cidades (Oakland, Richmond, New Haven, Chicago) nos EUA e abrangem um período de 12 meses (01/2012 - 12/2012). A versão do conjunto de dados que usamos aqui é uma amostra aleatória de 1% dos dados. Em termos de características, utilizamos o número de dias que um problema permaneceu online, a fonte de onde o problema foi criado (por exemplo, android, iphone, api remota, etc.), o tipo do problema (por exemplo, grafite, buraco na estrada, lixo, etc.), as coordenadas geográficas do problema, a cidade de onde foi publicado e a distância do centro da cidade. Todas as variáveis nominais de múltiplos valores foram primeiro transformadas em binárias e depois as variáveis binárias raras (verdadeiras para menos de 1% dos casos) foram removidas.

Slump

O conjunto de dados Concrete Slump (YEH, 2007) diz respeito à predição de 3 propriedades do concreto (abatimento, fluxo e resistência à compressão) como uma função com 7 variáveis de entrada relacionadas à ingredientes do concreto: cimento, cinzas volantes, escória de alto-forno, água, super plastificante, agregado graúdo e agregado miúdo.

Wq

O conjunto de Dados Water Quality (DŽEROSKI; DEMŠAR; GRBOVIĆ, 2000) tem 14 variáveis alvo que se referem a representação relativa de plantas e animais em rios eslovênicos e 16 variáveis de entrada que referem aos parâmetros físicos e químicos da água.

REFERÊNCIAS

AHO, T. et al. Multi-target regression with rule ensembles. **Journal of Machine Learning Research**, v. 13, n. 8, 2012.

ARGYRIOU, A.; EVGENIOU, T.; PONTIL, M. Convex multi-task feature learning. **Machine Learning**, v. 73, n. 3, p. 243–272, 2008.

BEN-DAVID, S.; SCHULLER, R. Exploiting task relatedness for multiple task learning. In: SCHÖLKOPF, B.; WARMUTH, M. K. (Ed.). **Learning Theory and Kernel Machines**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003. p. 567–580. ISBN 978-3-540-45167-9.

BORCHANI, H. et al. A survey on multi-output regression. **WIREs Data Mining and Knowledge Discovery**, v. 5, n. 5, p. 216–233, 2015. Disponível em: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1157.

CHUDZICKI, O. P. S. H. D. **Online Product Sales**. Kaggle, 2012. Online; acessado em 11 Junho 2024. Disponível em: https://kaggle.com/competitions/online-sales.

CUKIERSKI, W. **See Click Predict Fix**. Kaggle, 2013. Online; acessado em 11 Junho 2024. Disponível em: https://kaggle.com/competitions/see-click-predict-fix.

DŽEROSKI, S.; DEMŠAR, D.; GRBOVIĆ, J. Predicting chemical parameters of river water quality from bioindicator data. **Applied Intelligence**, v. 13, n. 1, p. 7–17, 2000. Disponível em: https://doi.org/10.1023/A:1008323212047.

GATTO, E. C.; FERRANDIN, M.; CERRI, R. Multi-label classification with label clusters. **PREPRINT (Version 1)**, 2023.

GOOVAERTS, P. **Geostatistics for Natural Resources Evaluation**. Oxford University Press, 1997. (Applied geostatistics series). ISBN 9780195115383. Disponível em: https://books.google.com.br/books?id=CW-7tHAaVR0C.

GROVES, W.; GINI, M. On optimizing airline ticket purchase timing. **ACM Transactions on Intelligent Systems and Technology**, Association for Computing Machinery (ACM), v. 7, n. 1, set. 2015. ISSN 2157-6904. Publisher Copyright: © 2015 ACM.

HADAVANDI, E.; SHAHRABI, J.; SHAMSHIRBAND, S. A novel boosted-neural network ensemble for modeling multi-target regression problems. **Engineering Applications of Artificial Intelligence**, v. 45, p. 204–219, 2015. ISSN 0952-1976. Disponível em: https://www.sciencedirect.com/science/article/pii/S095219761500144X.

HATZIKOS, E. et al. An empirical study on sea water quality prediction. **Knowledge-Based Systems**, v. 21, p. 471–478, 08 2008.

JEONG, M.; LEE, G. G. Multi-domain spoken language understanding with transfer learning. **Speech Communication**, v. 51, n. 5, p. 412–424, 2009. ISSN 0167-6393. Disponível em: https://www.sciencedirect.com/science/article/pii/S0167639309000028.

KARALIČ, A.; BRATKO, I. First order regression. **Machine Learning**, v. 26, n. 2, p. 147–176, 1997. Disponível em: https://doi.org/10.1023/A:1007365207130.

KOCEV, D.; CECI, M. Ensembles of extremely randomized trees for multi-target regression. In: JAPKOWICZ, N.; MATWIN, S. (Ed.). **Discovery Science**. Cham: Springer International Publishing, 2015. p. 86–100. ISBN 978-3-319-24282-8.

KOCEV, D. et al. Using single- and multi-target regression trees and ensembles to model a compound index of vegetation condition. **Ecological Modelling**, v. 220, n. 8, p. 1159–1168, 2009. ISSN 0304-3800. Disponível em: https://www.sciencedirect.com/science/article/pii/S0304380009000775.

KOCEV, D. et al. Tree ensembles for predicting structured outputs. **Pattern Recognition**, v. 46, n. 3, p. 817–833, 2013. ISSN 0031-3203. Disponível em: https://www.sciencedirect.com/science/article/pii/S003132031200430X.

KRIEGEL, H.-P. et al. Future trends in data mining. **Data Mining and Knowledge Discovery**, v. 15, n. 1, p. 87–97, 2007. Disponível em: https://doi.org/10.1007/s10618-007-0067-9.

MASMOUDI, S. et al. A machine-learning framework for predicting multiple air pollutants' concentrations via multi-target regression and feature selection. **Science of The Total Environment**, v. 715, p. 136991, 2020. ISSN 0048-9697. Disponível em: https://www.sciencedirect.com/science/article/pii/S0048969720305015.

MELKI, G. et al. Multi-target support vector regression via correlation regressor chains. **Information Sciences**, v. 415-416, p. 53–69, 2017. ISSN 0020-0255. Disponível em: https://www.sciencedirect.com/science/article/pii/S0020025517307946.

PEREIRA, R. B. et al. Categorizing feature selection methods for multi-label classification. **Artificial Intelligence Review**, v. 49, n. 1, p. 57–78, 2018. Disponível em: https://doi.org/10.1007/s10462-016-9516-4.

Rahimzadeh Arashloo, S.; KITTLER, J. Multi-target regression via non-linear output structure learning. **Neurocomputing**, v. 492, p. 572–580, 2022. ISSN 0925-2312. Disponível em: https://www.sciencedirect.com/science/article/pii/S092523122101883X.

SPYROMITROS-XIOUFIS, E. et al. Multi-target regression via input space expansion: treating targets as inputs. **Machine Learning**, v. 104, n. 1, p. 55–98, 2016. Disponível em: https://doi.org/10.1007/s10994-016-5546-z.

TSANAS, A.; XIFARA, A. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. **Energy and Buildings**, v. 49, p. 560–567, 2012. ISSN 0378-7788. Disponível em: https://www.sciencedirect.com/science/article/pii/S037877881200151X.

TSOUMAKAS, G. et al. Multi-target regression via random linear target combinations. In: CALDERS, T. et al. (Ed.). **Machine Learning and Knowledge Discovery in Databases**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014. p. 225–240. ISBN 978-3-662-44845-8.

TSOUMAKAS, G.; VLAHAVAS, I. Random k-labelsets: An ensemble method for multilabel classification. In: KOK, J. N. et al. (Ed.). **Machine Learning: ECML 2007**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. p. 406–417. ISBN 978-3-540-74958-5.

WANG, J. et al. Multi-target regression via target specific features. **Knowledge-Based Systems**, v. 170, p. 70–78, 2019.

YAN, Y. et al. A multi-task learning framework for head pose estimation under target motion. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 38, n. 6, p. 1070–1083, 2016.

YANG, Q.; WU, X. 10 challenging problems in data mining research. **International Journal of Information Technology and Decision Making (IJITDM)**, v. 05, p. 597–604, 12 2006.

YEH, I.-C. Modeling slump flow of concrete using second-order regressions and artificial neural networks. **Cement and Concrete Composites**, v. 29, n. 6, p. 474–480, 2007. ISSN 0958-9465. Disponível em: https://www.sciencedirect.com/science/article/pii/S0958946507000261.

YUAN, H. et al. Sparse structural feature selection for multitarget regression. **Knowledge-Based Systems**, v. 160, p. 200–209, 2018. ISSN 0950-7051. Disponível em: https://www.sciencedirect.com/science/article/pii/S0950705118303319.

ZHANG, M.-L.; ZHOU, Z.-H. A review on multi-label learning algorithms. **IEEE Transactions on Knowledge and Data Engineering**, v. 26, n. 8, p. 1819–1837, 2014.

ZHANG, Y.; YANG, Q. A survey on multi-task learning. **IEEE Transactions on Knowledge and Data Engineering**, v. 34, p. 5586–5609, 2017. Disponível em: https://api.semanticscholar.org/CorpusID:11311635.

ZHANG, Y.; YEUNG, D.-Y. **A Convex Formulation for Learning Task Relationships in Multi-Task Learning**. 2012.

ZHEN, X. et al. Multi-scale deep networks and regression forests for direct bi-ventricular volume estimation. **Medical Image Analysis**, v. 30, p. 120–129, 2016. ISSN 1361-8415. Disponível em: https://www.sciencedirect.com/science/article/pii/S1361841515001024.

ZHEN, X. et al. Multi-target regression via robust low-rank learning. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 40, n. 2, p. 497–504, 2018.