

## **ANÁLISE DO ESPAÇO DE BUSCA DE PARÂMETROS PARA O SUPPORT VECTOR REGRESSION (SVR)**

**Monica Beltrami**

Instituto Federal do Paraná - IFPR  
Rua João Negrão, 1285 – Rebouças, Curitiba – PR, 80230-150  
monica.beltrami@ifpr.edu.br

**Arinei Carlos Lindbeck da Silva**

Universidade Federal do Paraná – UFPR  
Programa de Pós-Graduação em Engenharia da Produção - PPGE  
Av. Cel. Francisco H. dos Santos, 100 - Jardim das Américas, Curitiba - PR, 81530-000  
arineicls@gmail.com

### **RESUMO**

O *Support Vector Regression* (SVR) tem apresentado excelentes resultados na área da regressão de dados, embora sua eficácia dependa diretamente da escolha de seus parâmetros. Diante disso, esse trabalho analisou as características do espaço de busca de parâmetros do SVR, utilizando-se o kernel gaussiano. Três bases de dados foram estudadas usando a busca por grid. Pelos resultados, constatou-se a existência de uma boa região de parâmetros, onde as soluções ótimas foram encontradas. Essas soluções propiciaram bom desempenho ao SVR, predizendo adequadamente os dados de teste com alto coeficiente de correlação. Embora sejam necessárias análises mais aprofundadas sobre o tema, identificar a boa região abre caminhos para novas pesquisas, em especial para o desenvolvimento de métodos mais eficazes de seleção de parâmetros, que se concentrem nessa zona específica. Isso otimiza a busca e reduz o tempo de processamento, ao evitar a avaliação de parâmetros em regiões de *underfitting* e *overfitting*.

**PALAVRAS CHAVE.** Seleção de parâmetros, Support Vector Regression, Espaço de busca.

**Tópicos.** EST&AM – PO Analytics em Estatística e Aprendizado de Máquina, OA – Outras Aplicações em PO

### **ABSTRACT**

Support Vector Regression (SVR) is a powerful method that has been successfully applied to many regression problems. However, the effectiveness of SVR depends on the proper selection of its parameters. For this reason, this paper analyzed the characteristics of the SVR parameter search space for the Gaussian kernel function. Three datasets were studied using grid search. The experimental results revealed the existence of a good region where optimal parameter solutions could be identified. These solutions enabled SVR to achieve good performance, accurately predicting test data with a high correlation coefficient. Although further in-depth analyses are needed, identifying this optimal region contributes to new research, especially in developing more effective parameter selection methods that focus on this specific zone. Such approaches can optimize the search process and reduce processing time by avoiding the evaluation of parameters in regions prone to underfitting or overfitting.

**KEYWORDS.** Parameter selection. Support Vector Regression. Search space.

**Paper topics.** EST&AM – OR Analytics in Statistics and Machine Learning, OA – Other Applications in OR

## 1. Introdução

O *Support Vector Regression* (SVR) é uma extensão do algoritmo *Support Vector Classification* (SVC), desenvolvido por [Boser et al. 1992], e destinado à regressão de dados. Devido à sua alta capacidade de generalização e excelentes resultados, o SVR tem sido amplamente aplicado em diversas áreas do conhecimento, como precificação de ações [Beniwal et al. 2023] e do crédito de carbono [Lu et al. 2025], previsão do nível de água em rios [Velasco et al. 2024], prognóstico de carga elétrica de curto prazo [Yang et al. 2019] e avaliação de tensão residual em soldas [Yazir et al. 2022].

A performance do SVR depende da correta seleção de seus parâmetros: constante de regularização  $C$ , função kernel e margem  $\varepsilon$ . Para auxiliar os usuários nessa tarefa, diversos métodos de seleção de parâmetros têm sido desenvolvidos. Entretanto, o método clássico, denominado busca por grid, continua sendo o mais utilizado devido à sua simplicidade.

A busca por grid é uma técnica exaustiva que avalia todas as combinações de parâmetros dentro de seu espaço de busca. Embora seja de fácil aplicação, apresenta um alto custo computacional. Por outro lado, ela possibilita, por meio da expressão de seus resultados, a compreensão das características do espaço formado pelos parâmetros e da relação entre eles. Esse entendimento é essencial para delimitar a amplitude do espaço de busca e para o desenvolvimento de métodos de seleção de parâmetros que priorizem a boa região, conforme será mostrado ao longo deste artigo.

Assim, esse trabalho tem como objetivo analisar o espaço de busca dos parâmetros do SVR – formado pela constante de regularização  $C$ , parâmetro  $\gamma$  do kernel gaussiano e margem  $\varepsilon$  – e seus efeitos na performance do algoritmo. Essa análise possibilita uma interpretação mais estruturada das relações entre os parâmetros, permitindo a definição de uma região de busca mais eficiente. Para isso, foram estudadas três bases de dados referência disponibilizadas na biblioteca LIBSVM de [Chang e Lin 2011].

## 2. Support Vector Regression (SVR)

O SVR emprega a metodologia do aprendizado supervisionado, para que o algoritmo aprenda uma função a partir de um conjunto de dados de treinamento  $\{\mathbf{x}_i, y_i\}_{i=1}^l$ . O objetivo do SVR é encontrar uma função  $f(\mathbf{x})$  que apresente no máximo um desvio  $\varepsilon$  em relação aos valores alvo  $y_i$ .

De acordo com [Smola e Schölkopf 2004], procura-se uma função que até pode admitir erros, desde que esses não ultrapassem a margem especificada pelo intervalo  $[y_i - \varepsilon, y_i + \varepsilon]$ . Além disso, espera-se que ela seja a mais paralela do domínio possível. Assumindo uma função na forma  $f(\mathbf{x}) = (\boldsymbol{\omega} \cdot \mathbf{x}) + b$ , isso significa determinar valores pequenos para o vetor dos pesos  $\boldsymbol{\omega}$ .

No entanto, garantir a existência de uma função  $f(\mathbf{x})$  que aproxime todos os pares  $(\mathbf{x}_i, y_i)$  com uma precisão  $\varepsilon$ , especificada a priori, nem sempre é viável. Por essa razão, é preciso considerar uma função de perda, que introduza variáveis de folga não negativas  $\xi_i, \xi_i^*$ , com a finalidade de penalizar os dados localizados fora da margem  $|f(\mathbf{x}_i) - y_i| \leq \varepsilon$ . A referida função é chamada de  $\varepsilon$ -Insensitive e descrita por:

$$\begin{aligned} |\xi|_\varepsilon &= 0, & \text{se } |\xi| &\leq \varepsilon \\ |\xi|_\varepsilon &= |\xi| - \varepsilon & \text{caso contrário.} \end{aligned} \quad (1)$$

A Figura 1 ilustra graficamente as condições propostas por (1), evidenciando que somente os pontos localizados fora da região sombreada contribuem para o valor custo da função.

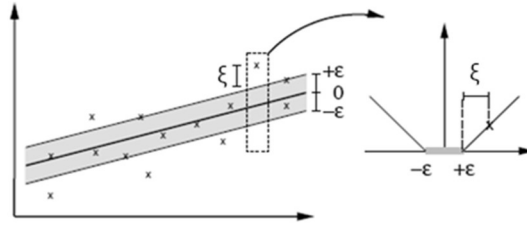


Figura 1 – Função de perda  $\varepsilon$ -Insensitive [Smola e Schölkopf 2004].

Com base no exposto, o modelo primal do SVR é formulado por

$$\begin{aligned}
 & \text{minimizar} \quad \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^l (\xi_i + \xi_i^*) \\
 & \text{sujeito a:} \quad y_i - \langle \omega \cdot x_i \rangle - b \leq \varepsilon + \xi_i \\
 & \quad \quad \quad \langle \omega \cdot x_i \rangle + b - y_i \leq \varepsilon + \xi_i^* \\
 & \quad \quad \quad \xi_i, \xi_i^* \geq 0, \quad \forall i = 1 \dots l
 \end{aligned} \tag{2}$$

onde  $\omega \in \mathbb{R}^n$ ,  $b \in \mathbb{R}$  são as incógnitas do problema.

No modelo (2),  $C$  é denominada constante de regularização, pois pondera os termos da função objetivo. Essa constante, que é um dos parâmetros do SVR, designa o grau de importância atribuído às variáveis de folga, permitindo uma maior ou menor quantidade de erros (desvios acima de  $\varepsilon$ ).

Em virtude das restrições de desigualdade, o problema de otimização (2) é mais facilmente resolvido por meio de sua formulação dual (3), que é obtida a partir de conceitos da teoria Lagrangeana. O modelo dual, além de facilitar a solução do SVR, propicia grande flexibilidade e eficiência ao algoritmo. Ela permite, por meio do emprego de funções kernel  $K(x_i \cdot x_j)$ , a realização de um mapeamento não linear dos dados de entrada para o espaço de características, de dimensão superior, onde a regressão linear torna-se possível [Beltrami et al. 2010]. Assim, o modelo dual do SVR é dado por

$$\begin{aligned}
 & \text{maximizar} \quad -\frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(x_i \cdot x_j) - \varepsilon \sum_{i=1}^l (\alpha_i + \alpha_i^*) + \\
 & \quad \quad \quad \sum_{i=1}^l y_i (\alpha_i - \alpha_i^*) \\
 & \text{sujeito a:} \quad \sum_{i=1}^l (\alpha_i - \alpha_i^*) = 0 \\
 & \quad \quad \quad \alpha_i, \alpha_i^* \in [0, C], \quad \forall i = 1 \dots l
 \end{aligned} \tag{3}$$

onde  $\alpha_i, \alpha_i^*$  são os multiplicadores de Lagrange.

Para o modelo (3), das condições de complementaridade de Karush-Kuhn-Tucker (KKT), tem-se que:

$$\alpha_i (\varepsilon + \xi_i - y_i + \langle \omega \cdot x_i \rangle + b) = 0, \quad \forall i = 1 \dots l \tag{4}$$

$$\alpha_i^* (\varepsilon + \xi_i^* + y_i - \langle \omega \cdot x_i \rangle - b) = 0, \quad \forall i = 1 \dots l \tag{5}$$

$$(C - \alpha_i) \xi_i = 0, \quad \forall i = 1 \dots l \tag{6}$$

$$(C - \alpha_i^*) \xi_i^* = 0, \quad \forall i = 1 \dots l \tag{7}$$

$$\alpha_i \alpha_i^* = 0, \quad \forall i = 1 \dots l \tag{8}$$

$$\xi_i \xi_i^* = 0, \quad \forall i = 1 \dots l \tag{9}$$

Avaliando a equação (8), constata-se que nunca existirá um conjunto de variáveis duais  $\alpha_i$  e  $\alpha_i^*$ , em que ambas serão não nulas. Das relações (4) e (5), observa-se que somente os dados que satisfazem  $|f(x_i) - y_i| \geq \varepsilon$  estão associados a multiplicadores de Lagrange diferente de zero. Tais dados estão localizados sobre as margens  $\varepsilon$  e  $-\varepsilon$  ou fora da região delimitadas por elas, e são os únicos que participam do cálculo de  $\omega$ , sendo chamados, portanto, de vetores suporte (VS).

Os vetores suportes podem ser limitados ( $\alpha_i = C$  ou  $\alpha_i^* = C$ ), também conhecidos por *VS Bound*, ou não limitados ( $0 < \alpha_i < C$  ou  $0 < \alpha_i^* < C$ ). Pelas condições (6) e (7), verifica-se que os pertencentes ao primeiro grupo possuem sua respectiva variável de folga  $\xi_i$  e  $\xi_i^*$  não nulas. Por consequência, localizam-se fora das margens e correspondem aos erros do modelo. Já os do segundo grupo, situam-se sobre as margens da função.

### 3. Influência da seleção de parâmetros para o SVR

A capacidade de generalização de um sistema de aprendizagem consiste na sua habilidade de prever adequadamente os dados de teste. Essa aptidão está diretamente relacionada à forma como se executa a etapa de treinamento, por meio da qual pode-se obter funções que melhor ou pior se ajustam as informações de entrada.

No caso do SVR, o desempenho do algoritmo está fortemente vinculado à seleção de seus parâmetros. O uso de diferentes funções kernel possibilita a construção de distintas funções de decisão, cuja complexidade e quantidade de erros é determinada pelo valor da constante de regularização  $C$  e da margem  $\varepsilon$ . A seleção inadequada dessas variáveis pode levar a fenômenos indesejáveis como o *overfitting* e o *underfitting*, que afetam tanto a acurácia da regressão quanto o tempo do treinamento [Wang et al. 2014].

A ocorrência do *overfitting* consiste no ajuste excessivo do algoritmo aos dados de treinamento. Em outras palavras, encontra-se uma função de decisão muito complexa e flexível, adaptada inclusive aos ruídos dos padrões de entrada. Assim, por se especializar demasiadamente no conjunto de treinamento, o algoritmo não consegue prever corretamente as saídas dos dados de teste, perdendo sua capacidade de generalização. São características do *overfitting*: alta taxa de acertos na fase de treinamento e baixa acurácia na etapa de teste.

Em contrapartida, o *underfitting* acontece quando se determina uma função muito simples e rígida, cuja flexibilidade é insuficiente para capturar as informações importantes dos padrões de entrada. Desta forma, o modelo se ajusta muito pouco aos dados de treinamento, não os representando adequadamente. Assim, no *underfitting*, tanto a acurácia do treinamento quanto a de teste são baixas.

No SVR, a ocorrência do *overfitting* é mais facilmente perceptível que a do *underfitting*, pois ela impacta diretamente na quantidade de vetores suporte do problema. Desta forma, o número de vetores suporte empregado na construção da função de decisão fornece informações importantes quanto a qualidade da função. Quantidades elevadas de vetores suporte, que se aproximam ou se igualam a totalidade do conjunto de treinamento, indicam um sobre ajuste do sistema (*overfitting*), enquanto que um baixo número de vetores suporte reflete numa melhor generalização dos dados. Nesse contexto, [Wang et al. 2023] destacam a importância da correta determinação do valor margem  $\varepsilon$ , uma vez que esse parâmetro controla o número de vetores suporte da função.

Em relação aos tipos de kernel, os mais comumente empregados no SVR são as funções: linear, polinomial homogênea, polinomial não homogênea, sigmoidal e gaussiana. Esta última, também denominada função base radial (RBF), pode ser representada pelas equações (10) e (11).

$$K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}} \quad (10)$$

$$K(x_i, x_j) = e^{-\gamma\|x_i - x_j\|^2} \quad (11)$$

Desta forma, o parâmetro da função RBF pode ser determinado por  $\sigma^2$  ou  $\gamma$ , dependendo da equação utilizada. No entanto, nota-se que (10) e (11) são essencialmente iguais, já que

$$\gamma = \frac{1}{2\sigma^2}. \quad (12)$$

Em virtude do seu amplo domínio de convergência e vasta gama de aplicações, a RBF é a função mais utilizada como kernel no SVR [Pang et al. 2011]. Por essa razão, quando se trata da seleção de parâmetros do SVR, a preocupação dos usuários consiste em determinar valores adequados para a constante de regularização  $C$ , margem  $\varepsilon$  e para o parâmetro  $\sigma^2$  ou  $\gamma$  do kernel gaussiano.

A relação entre os parâmetros  $C$  e  $\sigma^2$ , assim como as características do seu espaço de busca são conhecidas graças ao trabalho de [Keerthi e Lin 2003]. Esses autores analisaram o comportamento assintótico do *Support Vector Classification* (SVC) para os casos em que  $C$  e  $\sigma^2$  assumem valores muito grandes ou bem pequenos, conforme mostra a Figura 2.

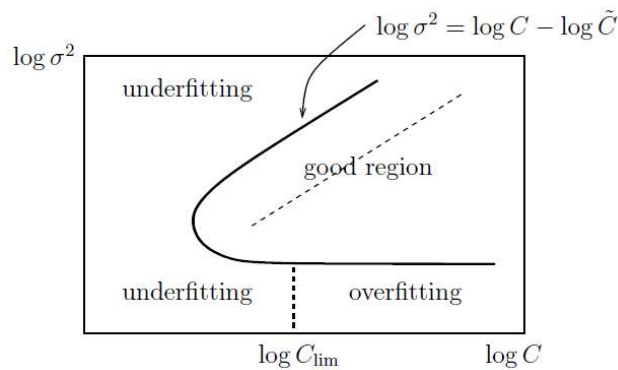


Figura 2 – Espaço de busca dos parâmetros  $C$  e  $\sigma^2$  [Keerthi e Lin 2003].

A partir da Figura 2, constata-se a existência de uma curva de erro de generalização que separa a área de busca dos parâmetros em duas regiões: uma caracterizada pelo *underfitting* e pelo *overfitting*, e outra denominada boa região (*good region*). Nessa última, encontra-se o par de parâmetros ótimo, responsável pela construção da função de decisão com melhor capacidade de generalização. Logo, essa é a região de interesse na seleção de parâmetros para o SVC.

Contudo, ao trabalhar com o algoritmo SVR, além de  $C$  e  $\sigma^2$ , é necessário determinar o valor da margem  $\varepsilon$ . No entanto, não foram encontradas na literatura informações que expliquem as características da região de busca dos parâmetros ( $C$ ,  $\sigma^2$ ,  $\varepsilon$ ), o que torna essa análise essencial.

#### 4. Busca por grid

A busca por grid (BG) é um método exaustivo que tem como finalidade encontrar parâmetros ótimos em uma malha ou grade. De acordo com [Garcia et al. 2009], a BG foi originalmente desenvolvida para selecionar parâmetros para problemas de classificação, ou seja, para o SVC. Portanto, para atender ao SVR, é necessário adaptar a faixa de busca para a margem  $\varepsilon$ . Consequentemente, por envolver a seleção de três parâmetros ( $C$ ,  $\sigma^2$ ,  $\varepsilon$ ), a malha assume uma estrutura cúbica.

A fim de aprimorar a habilidade de generalização do SVR, a BG emprega, em seu processo de investigação, a metodologia de validação cruzada do tipo *k-fold*. Segundo [Hsu et al. 2010], esse procedimento visa prevenir a ocorrência de *overfitting*. Assim, para cada ponto ( $C$ ,  $\sigma^2$ ,  $\varepsilon$ ) do espaço de busca, a BG efetua uma operação de validação cruzada, que consiste em dividir o conjunto de treinamento em  $k$  subconjuntos do mesmo tamanho, treinar o algoritmo com  $k - 1$  deles e testá-lo com o remanescente. Esse processo é repetido  $k$  vezes, garantindo que todos os subconjuntos sejam avaliados. Ao final das  $k$  iterações, calcula-se a precisão do modelo, que, no caso do SVR, consiste no erro quadrático médio e o coeficiente de correlação quadrático da validação cruzada.

Ressalta-se que o tamanho do espaço de busca, determinado pela faixa de procura dos parâmetros e pela espessura da malha, influencia diretamente no tempo computacional despendido pela BG, assim como a dimensão do conjunto de treinamento (número de dados e de características) e o valor do  $k$  da validação cruzada  $k$ -fold.

## 5. Metodologia

Para investigar a relação entre os parâmetros ( $C$ ,  $\gamma$ ,  $\epsilon$ ) do SVR, três grids cúbicos foram processados e analisados. O objetivo é verificar graficamente se, no espaço tridimensional, os parâmetros apresentam comportamento semelhante ao da Figura 2, possibilitando a obtenção de uma superfície que separa boa região das zonas de *underfitting* e *overfitting*.

Para isso, a BG foi implementada por meio da linguagem de programação VB.Net, e os cálculos referentes ao SVR foram efetuados usando pacotes computacionais disponibilizados na biblioteca LIBSVM [Chang e Lin 2011]. Já os desenhos gráficos dos grids cúbicos foram gerados utilizando a linguagem Python.

As bases de dados utilizadas neste estudo foram: *Bodyfat*, *Housing* e *Eunite2001*, disponíveis na LIBSVM. As duas primeiras bases, por não estarem inicialmente separadas em conjuntos de treinamento e teste, foram divididas aleatoriamente na proporção de 80% dos dados para treinamento e 20% para teste. Para essas duas, foram escolhidos os arquivos com os dados normalizados no intervalo  $[-1,1]$ . Quanto à base *Eunite2001*, por ela estar originalmente dividida em conjuntos de treinamento e teste, seus dados foram utilizados tal como foram disponibilizados na plataforma. O Quadro 1 descreve o número de características e a quantidade de dados de cada um dos conjuntos analisados.

Quadro 1 – Conjunto de dados avaliados.

| Base de dados     | Conjunto    | Quantidade de dados | Número de características |
|-------------------|-------------|---------------------|---------------------------|
| <i>Bodyfat</i>    | Treinamento | 202                 | 14                        |
|                   | Teste       | 50                  |                           |
| <i>Housing</i>    | Treinamento | 405                 | 13                        |
|                   | Teste       | 101                 |                           |
| <i>Eunite2001</i> | Treinamento | 336                 | 16                        |
|                   | Teste       | 31                  |                           |

Nos três casos avaliados, verificaram-se as características do espaço de parâmetros em uma malha  $17 \times 17 \times 17$ , onde o eixo da constante  $C$  variou no intervalo de  $[2^{-5}, 2^{-4}, \dots, 2^{10}, 2^{11}]$  e o eixo de  $\gamma$  entre  $[2^{-8}, 2^{-7}, \dots, 2^7, 2^8]$ . Uma vez que a margem  $\epsilon$  depende dos dados analisados, seu intervalo não pode ser padronizado como os de  $C$  e  $\gamma$ . Assim, o limite máximo de  $\epsilon$  foi calculado como o desvio padrão dos valores alvo  $y_i$  dos dados de treinamento. Além disso, como  $\epsilon > 0$ , os respectivos valores dos desvios-padrão ( $\sigma_{y_i}$ ) foram divididos em 17 partes, sendo o valor mínimo do intervalo igual ao valor dessa divisão. O Quadro 2 apresenta os valores dos intervalos da margem  $\epsilon$  empregadas para cada base de dados.

Quadro 2 –Intervalo da margem  $\epsilon$  para os conjuntos analisados.

| Base de dados     | Desvio-padrão dos valores alvo ( $\sigma_{y_i}$ ) | Intervalo da margem $\epsilon$              |
|-------------------|---------------------------------------------------|---------------------------------------------|
| <i>Bodyfat</i>    | 0,01969335                                        | [0,001158432; 0,002316865; ...; 0,01969335] |
| <i>Housing</i>    | 9,342898                                          | [0,5495822; 1,099164; ...; 9,342898]        |
| <i>Eunite2001</i> | 47,885                                            | [2,816765; 5,633529; ...; 47,885]           |

Por fim, nas três situações analisadas, utilizou-se a técnica de validação  $k$ -fold com  $k=5$ .

## 6. Resultados

Para determinar o melhor trio de parâmetros ( $C$ ,  $\gamma$ ,  $\epsilon$ ) avaliado pela busca por grid, utilizaram-se como medidas de desempenho o percentual de vetores suporte da solução e o coeficiente de correlação quadrático da validação cruzada. Assim, as soluções deveriam apresentar uma quantidade de vetores suporte inferior a 50% e um coeficiente de correlação superior a 70%, de forma a evitar a seleção de parâmetros que levassem ao *overfitting*. Ressalta-se ainda que coeficientes de correlação quadráticos próximos de 100% indicam um bom ajuste do modelo, explicando bem a relação entre as variáveis.

A Tabela 1 apresenta as características das respostas encontradas.

Tabela 1 – Parâmetros ótimos.

| Solução                                                   | <i>Bodyfat</i>           | <i>Housing</i>  | <i>Eunite2001</i>  |
|-----------------------------------------------------------|--------------------------|-----------------|--------------------|
| $C$                                                       | $2^{-1} = 0,5$           | $2^7 = 128$     | $2^{11} = 2048$    |
| $\gamma$                                                  | $2^{-7} = 0,0078125$     | $2^{-2} = 0,25$ | $2^{-5} = 0,03125$ |
| $\epsilon$                                                | 0,001158432              | 1,099164        | 11,26706           |
| Coeficiente de correlação quadrático da validação cruzada | 0,971449                 | 0,888819        | 0,831195           |
| Erro quadrático médio da validação cruzada                | $1,11814 \times 10^{-5}$ | 9,74535         | 387,032            |
| VS                                                        | 38                       | 199             | 144                |
| VS (%)                                                    | 18,81%                   | 49,14%          | 42,86%             |
| VS <i>Bound</i>                                           | 15                       | 82              | 128                |

A Figura 3 ilustra três grids planos, onde os eixos X e Y representam, respectivamente, as faixas de busca dos parâmetros  $C$  e  $\gamma$ , enquanto o valor de  $\epsilon$  permanece constante e equivale ao seu valor ótimo. O grid à esquerda corresponde à base de dados *Bodyfat*, o do meio à *Housing* e o da direita à *Eunite2001*. Nas imagens, os pontos vermelhos indicam os trios de parâmetros ótimos, cujos valores constam na Tabela 1.

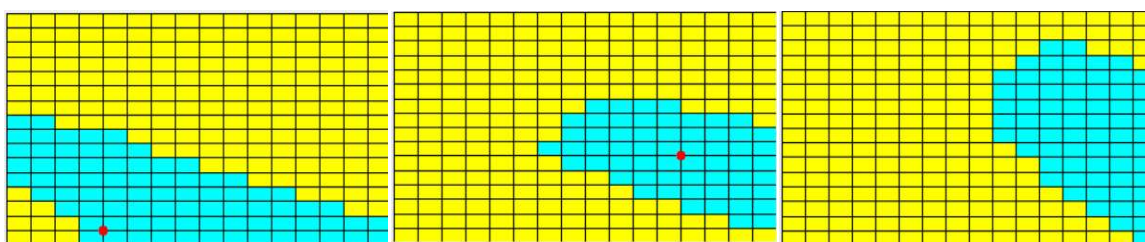


Figura 3 – Solução gráfica da busca por grid, no plano, para as três bases de dados.

Na Figura 3, as áreas em azul representam a boa região de parâmetros, que foram delineadas considerando apenas conjuntos de parâmetros com coeficiente de correlação superior a 0,70. Além dessa métrica, estabeleceu-se um percentual máximo para a quantidade de VS, com valores de 45%, 50% e 55%, respectivamente, para a *Bodyfat*, *Eunite2001* e *Housing*.

Pela figura 3, observa-se que, para as bases *Housing* e *Eunite2001*, o espaço de parâmetros exibiu um comportamento semelhante ao apresentado na Figura 2. Já para a *Bodyfat*, visualizam-se mais diferenças em relação a [Keerthi e Lin 2003], as quais podem ser explicadas pela escolha da faixa de valores do parâmetro  $\gamma$ . Ao determinar para o eixo  $\gamma$  valores entre  $[2^{-13}, 2^{-12}, \dots, 2^2, 2^3]$ , a separação da boa região de parâmetros das demais áreas torna-se mais nítida, conforme mostra a Figura 4.

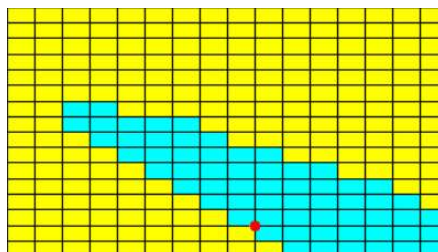


Figura 4 – Solução gráfica da busca por grid, no plano, para a base *Bodyfat* com nova faixa de  $\gamma$ .

Novamente, na Figura 4, o ponto vermelho indica o trio de parâmetros ótimos que, para essa nova configuração do grid, resultou em  $(C, \gamma, \epsilon) = (2^4, 2^{-11}, 0,001158432)$ . Destaca-se que o valor do parâmetro  $\epsilon$  permaneceu igual ao listado na Tabela 1. Contudo, os valores do MSE e do coeficiente de correlação da validação cruzada registraram uma ligeira melhora, atingindo  $1,07862 \times 10^{-5}$  e 0,972091, respectivamente. Da mesma forma, a quantidade total de VS foi reduzida para 15,35%. A delimitação da região azul da Figura 4 foi realizada considerando apenas parâmetros que apresentaram coeficiente de correlação superior a 0,9 e um percentual máximo de VS igual a 20%.

Essa mudança na faixa de  $\gamma$ , para a base *Bodyfat*, demonstra a importância de se delimitar adequadamente o espaço de busca de parâmetros. Além disso, evidencia que as melhores soluções tendem a se concentrar numa determinada região, uma vez que ambas as soluções (Figuras 3 e 4) estão muito próximas, permanecendo, inclusive, no mesmo plano de  $\epsilon$ .

Assim, para analisar a distribuição da boa região no espaço tridimensional, todas as combinações de parâmetros  $(C, \gamma, \epsilon)$  avaliadas pela busca por grid foram plotadas na Figura 5. Novamente, o grid à esquerda representa a base de dados *Bodyfat*, o do meio a *Housing* e o da direita a *Eunite2001*. As regiões azuis foram desenhadas considerando os mesmos valores das métricas de desempenho empregadas na Figura 3, assim como os eixos de  $C$  e  $\gamma$ .

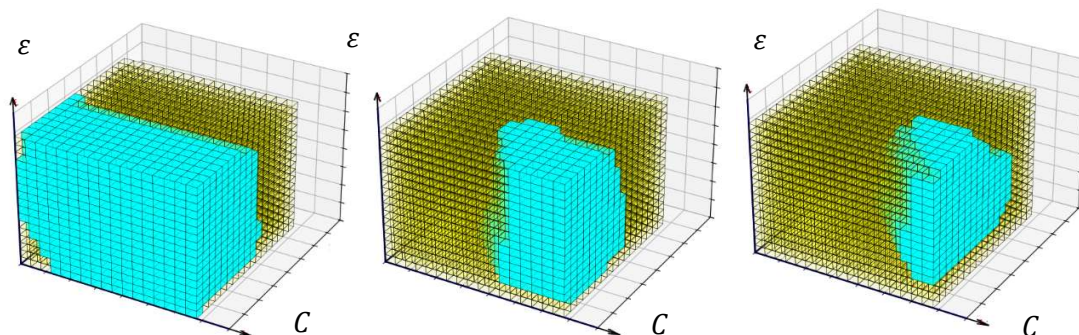


Figura 5 – Representação tridimensional do espaço de parâmetros para as três bases de dados.

Pela Figura 5, repara-se que, para a base *Bodyfat*, muitos parâmetros atendem aos critérios de coeficiente de correlação superior a 0,70 e percentual de vetores suporte inferior a 45%. No entanto, essa alta concentração de parâmetros dificulta a visualização das regiões desejadas e indesejadas no espaço tridimensional. Por essa razão, na Figura 6, a região azul foi limitada a soluções com coeficiente de correlação superior a 0,95 e VS inferior a 20%, tornando sua percepção mais clara.

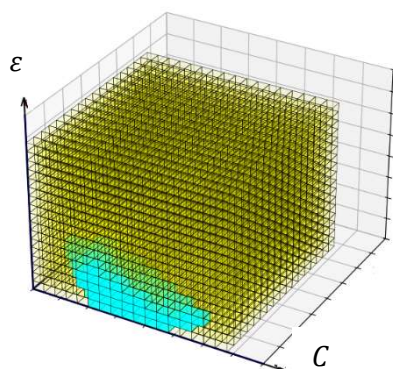


Figura 6 – Representação tridimensional do espaço de parâmetros da base *Bodyfat*.

Com o intuito de avaliar a eficiência dos parâmetros listados na Tabela 1 na performance do SVR, esses foram utilizados na previsão dos dados de teste. A Tabela 2 apresenta os resultados obtidos para as três bases de dados, em termos de erro quadrático médio (MSE) e coeficiente de correlação quadrático.

Tabela 2 – Parâmetros ótimos.

| Solução                              | <i>Bodyfat</i>            | <i>Housing</i> | <i>Eunite2001</i> |
|--------------------------------------|---------------------------|----------------|-------------------|
| MSE                                  | $2,128956 \times 10^{-6}$ | 13,4481        | 253,332           |
| Coeficiente de correlação quadrático | 0,991874                  | 0,822549       | 0,792809          |

Pelos valores da tabela 2, observa-se que o SVR conseguiu capturar bem a relação entre os dados, apresentando boa correlação entre os valores previstos e os reais, para as três bases de dados. Embora excelente para a base *Bodyfat*, o MSE obtido para a base *Eunite2001* sugere uma diferença considerável entre as estimativas do modelo e os valores originais. Contudo, o valor alcançado na etapa de teste (253,332) foi aproximadamente 34,5% menor do que o da validação cruzada (387,032) do treinamento.

Ainda, quando explorado um novo espaço de parâmetros para a base *Bodyfat* conforme Figura 4, o trio ótimo  $(C, \gamma, \epsilon) = (2^4, 2^{-11}, 0,001158432)$  apresentou um resultado ligeiramente melhor na previsão dos dados de teste do que o listado na Tabela 2. Para esses novos parâmetros, o MSE foi de  $2,04228 \times 10^{-6}$  e coeficiente de correlação igual a 0,99215.

Além disso, a seleção aleatória de algumas combinações de parâmetros em regiões indicadas como zonas de *underfitting* confirma a baixa acurácia tanto na etapa de treinamento quanto na de teste, para as três bases de dados, conforme mostra a Tabela 3. Esse comportamento é evidenciado principalmente pelos baixos coeficientes de correlação obtidos em ambas as etapas.

Tabela 3– Parâmetros localizados nas zonas de *underfitting*.

| Base de dados     | Parâmetros<br>( $C, \gamma, \epsilon$ ) | Treinamento              |                                                | Teste      |                           |
|-------------------|-----------------------------------------|--------------------------|------------------------------------------------|------------|---------------------------|
|                   |                                         | MSE da validação cruzada | Coeficiente de correlação da validação cruzada | MSE        | Coeficiente de correlação |
| <i>Bodyfat</i>    | $(2^{-5}, 2^7, 0,01853492)$             | 0,0003870                | 0,00787132                                     | 0,00025998 | $3,5755 \times 10^{-7}$   |
| <i>Housing</i>    | $(2^{-3}, 2^3, 4,396658)$               | 86,589                   | 0,0976687                                      | 71.4715    | 0.276358                  |
| <i>Eunite2001</i> | $(2^0, 2^{-6}, 5,633529)$               | 2307,36                  | 0,0143249                                      | 1235,27    | 0,302922                  |

## 7. Conclusão

Este trabalho conduziu uma pesquisa preliminar sobre as características do espaço de busca dos parâmetros do SVR, utilizando o kernel gaussiano. Nesse estudo, foram analisadas três bases de dados referência, disponibilizadas na biblioteca LIBSVM.

Paras essas bases, confirmou-se a existência de uma boa região de parâmetros, onde se encontraram as soluções ótimas. Essas combinações de parâmetros proporcionaram boa capacidade de generalização ao algoritmo SVR, com coeficientes de correlação superiores a 0,79, indicando uma alta correspondência entre os dados.

Embora o estudo necessite de investigações mais aprofundadas, constatar a existência de uma boa região de parâmetros possibilita o desenvolvimento de métodos de seleção de  $(C, \gamma, \epsilon)$ , que foquem nessa região específica. Isso otimiza o tempo de processamento da busca, evitando o teste de parâmetros localizados nas zonas de *underfitting* e *overfitting*, além de abrir campo para novas pesquisas.

Assim, pretende-se dar continuidade a esse estudo, ampliando a análise para novas bases de dados e expandindo o método *grid-quadtree* de [Beltrami e Silva 2020] para o SVR. Ademais, esse trabalho contribui com a definição de intervalos de busca para os parâmetros  $C$  e  $\gamma$ , em especial para a margem  $\epsilon$ , auxiliando a aplicação da busca por grid por usuários menos experientes.

## Referências

Beltrami M.; Silva, A. C. L. da (2020). A grid-quadtree model selection method for support vector machines *Expert Systems with Applications*, 146: 1-13.

Beltrami, M.; Loch, G. V e Silva, A. C. L. da (2010). Comparação das técnicas de support vector regression e redes neurais na precificação de opções. Artigo em conferência. In: *Anais do XLII SBPO*, p. 572-583, Bento Gonçalves. SOBRAPO.

Beniwal, M; Singh, A. e Kumar, N. (2023). Forecasting long-term stock prices of global índices: a forward validating genetic algorithm optimization approach for support vector regression. *Applied Soft Computing*, 145: 1-15.

Boser, B. E., Guyon, I. M. e Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. Artigo em conferência. In *Proceedings of the 5<sup>th</sup> Annual Workshop on Computational Learning Theory*, p. 144-152, Pittsburg.

Chang, C. C. e Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2: 1-27. Web-page. <http://www.csie.ntu.edu.tw/~cjlin/libsvm>. Acessado: 2025-05-07.

Garcia, E. G. O.; Sanz, S. S.; Bellido, A. M. P. e Figueras J. A. P. (2009). Improving the training time of support vector regression algorithms through novel hyper-parameters search space reductions. *Neurocomputing*, 72: 3363-3691.

Hsu, C. W.; Chang, C. C.; Lin, C. J. (2010). A practical guide to support vector classification. *Technical Report*. Department of Computer Science. National Taiwan University.

Keerthi, S. S. e Lin, C. J. (2003). Asymptotic behaviors of support vector machines with gaussian kernel. *Neural Computation*, 15: 1166-1189.

Lu, Y.; Wang, J. e Li Q. (2025). An interval-valued carbon price prediction model based on improved multi-scale feature selection and optimal multi-kernel support vector regression. *Information Sciences*, 692: 1-26.

Pang, H.; Dong, W. D.; Xu, Z. H.; Feng, H. J.; Li, Q. e Chen, Y. T. (2011). Novel linear search for support vector machine parameter selection. *Journal of Zhejiang University (Computers & Electronics)*, 12: 885-896.

Smola, A. J. e Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14: 199-222.

Velasco, L. C.; Estose, A. J.; Opon, M.; Tabanao, E. e Apdian, F. (2024). Performance evaluation of support vector regression machine models in water level forecasting. *Procedia Computers Science*, 234: 436-447.

Wang, X.; Huang, F.; Cheng, Y. (2014). Super-parameter selection for gaussian-kernel SVM based on outlier-resisting. *Measuremet*, 58: 147-153.

Wang, Y.G.; Wu, J.; Hu, Z. H. e McLahlan, G. J. (2023). A new algorithm for support vector regression with automatic selection of hyperparameters. *Pattern Recognition*, 133: 1-9.

Yang, Y.; Che, J.; Deng, C. e Li, L. (2019) Sequential grid approach based support vector regression for short-term electric load forecasting. *Applied Energy*, 238: 1010-1021.

Yazir, S. M; Dhas, J. E. R; Darwins, A. K.; Lewise, A. S. e Gupta, M. S. (2022). Modelling of weld residual stress parameters by SVR approach. *Materials Today: Proceeding*, 64: 338-344.