Feature Selection for Support Vector Regression Using Probabilistic Prediction

Jian-Bo Yang and Chong-Jin Ong

yangjianbo@nus.edu.sg
mpeongcj@nus.edu.sg

National University of Singapore
CONTENTS

Introduction

Proposed Method

Experiments

Conclusions
<table>
<thead>
<tr>
<th>Introduction</th>
<th>Proposed Method</th>
<th>Experiments</th>
<th>Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Introduction</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proposed Method</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experiments</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conclusions</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Background

Feature Selection is a technique of selecting optimal features set among original features set by removing irrelevant or redundant features.

**Benefits:**

- Increase system interpretability
- Improve generalization performance
- Minimize the overfitting for some learning algorithms

**Types:**

- *Filter Methods:* independent of the underlying learning algorithm
- *Wrapper Methods:* rely heavily on the specific structure of the underlying learning.

**Challenge:**

- Using feature selection for classification on regression problem may not work well — potential loss of important ordinal information.
Support Vector Regression

Given a data set $\mathcal{D} = \{x_i, y_i\}, i \in \mathcal{I}_D$, standard SVR solves the following Primal Problem (PP) over $\omega, b, \xi, \xi^*$:

$$
\min \quad \frac{1}{2} \omega'\omega + C \sum_{i \in \mathcal{I}_D} (\xi_i + \xi_i^*) \\
\text{s.t.} \quad y_i - \omega' \phi(x_i) - b \leq \epsilon + \xi_i, \quad \forall i \in \mathcal{I}_D \\
\omega' \phi(x_i) + b - y_i \leq \epsilon + \xi_i^*, \quad \forall i \in \mathcal{I}_D \\
\xi_i, \xi_i^* \geq 0, \quad \forall i \in \mathcal{I}_D
$$

The regressor function is known to be

$$
f(x) = \omega' \phi(x) + b
$$

It only provides an estimate, $f(x)$, for output $y$ for any $x$ but provides no information on the confidence level of this estimate.
A popular approach [Bishop 1995] to incorporating probabilistic information is to let

\[ y = f(x) + \delta. \]

where noise \( \delta \in \mathcal{L}(0, \sigma) \) or \( \mathcal{N}(0, \sigma) \)
Equivalently, this implies that density functions of \( y \) for a given \( x \) are

\[
p^L(y|x) = \frac{1}{2\sigma} \exp\left(-\frac{|y - f(x)|}{\sigma}\right),
\]

\[
p^G(y|x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(y - f(x))^2}{2\sigma^2}\right)
\]

where \( \sigma \) is obtained by maximizing

\[
L(\sigma) = \prod_{i \in \mathcal{D}} p(x_i, y_i) = \prod_{i \in \mathcal{D}} p(y_i|x_i)p(x_i).
\]
<table>
<thead>
<tr>
<th>Introduction</th>
<th>Proposed Method</th>
<th>Experiments</th>
<th>Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Introduction</strong></td>
<td><strong>Proposed Method</strong></td>
<td><strong>Experiments</strong></td>
<td><strong>Conclusions</strong></td>
</tr>
</tbody>
</table>

- **Proposed Method**

- **Experiments**

- **Conclusions**
Proposed Feature Selection Criterion

- **Ranking criterion:**

\[
S_D(j) = \int D_{KL}(p(y|x); p(y|x_j))p(x)dx.
\]

where \(x_j \in \mathbb{R}^{d-1}\) is the sample \(x\) with the \(j^{th}\) feature removed.

- **Motivation:**

the greater the \(D_{KL}\) divergence between \(p(y|x)\) and \(p(y|x_j)\) over the \(x\) space, the greater the importance of the \(j^{th}\) feature.

- A full ranking list of features need \(S_D(j)\) to be evaluated \(d\) times, each time with different \(j\).
Random Permutation

- **Random permutation:**

\[
D = \begin{pmatrix}
    x_1 \\
    x_2 \\
    \vdots \\
    x_N
\end{pmatrix} = \begin{pmatrix}
    x_1^1 & \ldots & x_j^j & \ldots & x_d^d \\
    x_1^1 & \ldots & x_j^j & \ldots & x_d^d \\
    \vdots & \vdots & \vdots & \vdots & \vdots \\
    x_1^N & \ldots & x_j^N & \ldots & x_d^N
\end{pmatrix}
\]

\[
\downarrow
\]

\[
D_{(j)} = \begin{pmatrix}
    x_1 \\
    x_2 \\
    \vdots \\
    x_N
\end{pmatrix} = \begin{pmatrix}
    x_1^1 & \ldots & x_j^j & \ldots & x_d^d \\
    x_1^1 & \ldots & x_j^j & \ldots & x_d^d \\
    \vdots & \vdots & \vdots & \vdots & \vdots \\
    x_1^N & \ldots & x_j^N & \ldots & x_d^N
\end{pmatrix}
\]

- **Theorem** [Shen, Ong, Li, & Wilder-Smith, 2008]: Assume data samples are sufficient rich,

\[
p(y|x_{(j)}) = p(y|x_{-j})
\]
Equivalent Form of the Proposed Criterion

\[ S_D(j) = \int D_{KL}(p(y|x); p(y|x_{(j)}))p(x)dx. \]

**Figure:** Demonstration of the proposed feature ranking criterion with \( d = 1 \). Dots indicate locations of \( y_i \).
Approximations

- Step 1: Further approximation of integration

\[ \hat{S}_D(j) = \frac{1}{|\mathcal{I}_D|} \sum_{i \in \mathcal{I}_D} D_{KL}(p(y|x_i); p(y|x(j),i)). \]

- Step 2: Approximation using probabilistic outputs of SVR

\[ \hat{S}_D(j) = \frac{1}{|\mathcal{I}_D|} \sum_{i \in \mathcal{I}_D} D_{KL}(p(y|x_i); p(y|x(j),i)). \]

\( p(.) \) can be approximated by \( p^L(.) \) or \( p^G(.) \)
Explicit form exist. E.g. if \( p(.) \) is approximated by \( p^L(.) \), then:

\[
\hat{S}_D^L(j) = \frac{1}{|\mathcal{I}_D|} \sum_{i \in \mathcal{I}_D} \left[ \frac{\sigma^L}{\sigma^L_{(j)}} \exp\left(-\frac{|f(x_i) - f(x_{(j),i})|}{\sigma^L_{(j)}}\right) + \frac{|f(x_i) - f(x_{(j),i})|}{\sigma^L_{(j)}} + \ln \frac{\sigma^L_{(j)}}{\sigma^L} \right].
\]

SD measure can be used together with standard recursive feature elimination (RFE).

1. Start with all features
2. Delete feature(s) with the smallest value(s) of \( \hat{S}_D^L \) (or \( \hat{S}_D^G \))
Introduction

Proposed Method

Experiments

Conclusions
Experiment Setting

- **Benchmark Methods:** Correlation coefficient method (Corr), Dependence maximization method (HSIC), SVM-RFE method ($\Delta \|\omega\|^2$)

- **Evaluation:** Mean squared error rate (MSE)

- **Student Test:**
  
  - Paired t-test between the proposed method and each of the other methods is conducted using different number of top ranked features.
  
  \[
  \begin{align*}
  \mu_0 : & \quad MSE_{SD} = MSE_{Benchmark} \\
  \mu_1 : & \quad MSE_{SD} \neq MSE_{Benchmark}
  \end{align*}
  \]

  The chance that this null hypothesis $\mu_0$ is true is measured by the returned p-value and the significance level is set at 0.05 for all experiments.
Artificial Problems

Table: Description of artificial problems. $o$ is the number of known important features.

| Problems           | $|D_{trn}|$       | $|D_{tst}|$ | $d$ | $o$ |
|--------------------|------------------|-------------|-----|-----|
| Exponential Func   | 100,70,50,40,30,20 | 1800        | 10  | 2   |
| Additive Func      | 200,100,70,50     | 1800        | 10  | 5   |
| Interactive Func   | 200,100,70,50     | 1800        | 10  | 5   |

Target Concept

- **Exponential Func:**
  \[
  y = 10 \exp(-((x^1)^2 + (x^2)^2)) + \delta
  \]

- **Additive Func:**
  \[
  y = 0.1 \exp(4x^1) + \frac{4}{1 + \exp(-20(x^2 - 0.5))} + 3x^3 + 2x^4 + x^5 + \delta
  \]

- **Interactive Func:**
  \[
  y = 10 \sin(\pi x^1 x^2) + 20(x^3 - 0.5) + 10x^4 + 5x^5 + \delta
  \]
### Table: Number of realizations that known important features are correctly ranked in the top positions over 30 realizations.

| Method \ $|\mathcal{D}_{trn}|$ | Exponential Func | Additive Func | Interactive Func |
|-------------------------------|------------------|---------------|------------------|
|                               | 100  70  50  40  30  20 | 200  100  70  50 | 200  100  70  50 |
| Corr                          | 0    0    0    0    0    0 | 15   8   5    3 | 4    3    2    1 |
| HSIC-RFE                      | 30   29   28   22   16   9 | 14   5   5    3 | 7    9    8    6 |
| $\Delta||\omega||^2$-RFE      | 30   30   28   28   1    0 | 4    5   11   4 | 0    14   9    10 |
| SD-L-RFE                      | 30   30   30   30   26   17 | 30   27   21   19 | 30   30   29   12 |
| SD-G-RFE                      | 30   30   29   28   26   13 | 30   28   23   19 | 30   30   30   11 |
Figure: Average test MSE against top-ranked features over 30 realizations.
### Real-World Problems

| Data sets   | \( |D_{trn}| \) | \( |D_{tst}| \) | \( d \) | \( C \) | \( \kappa \) | \( \epsilon \) |
|-------------|----------------|----------------|--------|-------|-------|--------|
| mpg         | 353            | 39             | 7      | \( 2^6 \) | \( 2^{-4} \) | 2      |
| abalone     | 1254           | 2923           | 8      | \( 2^6 \) | \( 2^{-5} \) | 2      |
| cpusmall    | 820            | 7372           | 12     | \( 2^6 \) | \( 2^{-5} \) | 2      |
| housing     | 456            | 50             | 13     | \( 2^6 \) | \( 2^{-4} \) | 2      |
| pyrim       | 67             | 7              | 27     | \( 2^0 \) | \( 2^{-6} \) | \( 2^{-5} \) |
| triazines   | 168            | 18             | 60     | \( 2^{-1} \) | \( 2^{-6} \) | \( 2^{-3} \) |

**Table:** Description of real-world data sets. \( C, \kappa \) and \( \epsilon \) refer to SVR hyper-parameters \( C, \kappa, \epsilon \) respectively.
**Table:** $t$-test on data set cpusmall for 30 realizations

<table>
<thead>
<tr>
<th>No.</th>
<th>SD-L-RFE</th>
<th>Corr</th>
<th>HSIC-RFE</th>
<th>$\Delta |\omega|^2$-RFE</th>
<th>SD-G-RFE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean value</td>
<td>mean value</td>
<td>p-value</td>
<td>mean value</td>
<td>p-value</td>
</tr>
<tr>
<td>2</td>
<td>40.39</td>
<td>74.38</td>
<td><strong>0.00+</strong></td>
<td>293.6</td>
<td><strong>0.00+</strong></td>
</tr>
<tr>
<td>4</td>
<td>18.99</td>
<td>27.66</td>
<td><strong>0.00+</strong></td>
<td>82.44</td>
<td><strong>0.00+</strong></td>
</tr>
<tr>
<td>6</td>
<td>19.20</td>
<td>22.33</td>
<td><strong>0.01+</strong></td>
<td>28.57</td>
<td>0.32</td>
</tr>
<tr>
<td>8</td>
<td>20.66</td>
<td>21.09</td>
<td>0.49</td>
<td>20.49</td>
<td>0.78</td>
</tr>
<tr>
<td>10</td>
<td>21.64</td>
<td>21.57</td>
<td>0.92</td>
<td>22.49</td>
<td>0.28</td>
</tr>
<tr>
<td>12</td>
<td>23.78</td>
<td>23.78</td>
<td>1.00</td>
<td>23.78</td>
<td>1.00</td>
</tr>
</tbody>
</table>
A new wrapper based feature selection method for regression problem is proposed. It measures the importance of a feature by the aggregation, over the feature space, of the sensitivity of SVR probabilistic prediction with and without the feature.

The experiments results show that the proposed method performs at least as well, if not better, than some of the benchmark methods in the literature.

The advantage of the proposed methods is more significant when the training data is sparse, or has a low samples-to-features ratio.

As a wrapper method, the computational cost of proposed methods is moderate.