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

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 $\mathcal{D} = \{x_i, y_i\}, i \in \mathcal{I}_\mathcal{D}$, standard SVR solves the following Primal Problem (PP) over ω, b, ξ, ξ^*:

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

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

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

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

where σ is obtained by maximizing

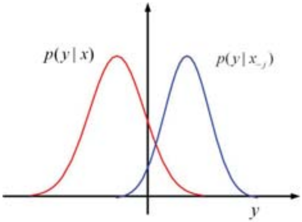
$$L(\sigma) = \Pi_{i \in \mathcal{I}_\mathcal{D}} p(x_i, y_i) = \Pi_{i \in \mathcal{I}_\mathcal{D}} p(y_i|x_i) p(x_i).$$

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**:

$$\mathbf{D} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} = \begin{pmatrix} x_1^1 & \dots & x_1^j & \dots & x_1^d \\ x_2^1 & \dots & x_2^j & \dots & x_2^d \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_N^1 & \dots & x_N^j & \dots & x_N^d \end{pmatrix}$$

$$\Downarrow$$

$$\mathbf{D_{(j)}} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} = \begin{pmatrix} x_1^1 & \dots & x_4^j & \dots & x_1^d \\ x_2^1 & \dots & x_1^j & \dots & x_2^d \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_N^1 & \dots & x_6^j & \dots & x_N^d \end{pmatrix}$$

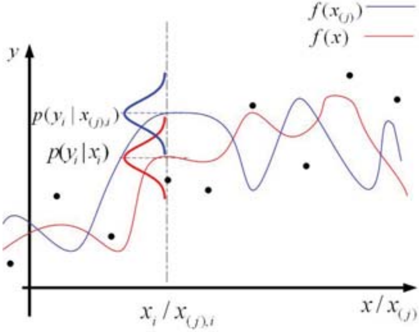
- **Theorem** [Shen, Ong, Li, & Wilder-Smith, 2008]: Assume data samples are sufficient rich,

$$p(y|x_{(j)}) = p(y|x_{-j})$$

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

$$\hat{S}_D(j) = \frac{1}{|\mathcal{I}_\mathcal{D}|} \sum_{i \in \mathcal{I}_\mathcal{D}} D_{KL}(p(y|x_i); p(y|x_{(j),i})).$$

- Step 2: Approximation using probabilistic outputs of SVR

$$\hat{S}_D(j) = \frac{1}{|\mathcal{I}_\mathcal{D}|} \sum_{i \in \mathcal{I}_\mathcal{D}} D_{KL}(p(y|x_i); p(y|x_{(j),i})).$$

$p(.)$ can be approximated by $p^L(.)$ or $p^G(.)$

Explicit form exist. E.g. if $p(.)$ is approximated by $p^L(.)$, then:

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

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` $\hat{S}_D^G$`)`

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.
  - $$\begin{aligned} \mu_0 : &\quad MSE_{SD} = MSE_{Benchmar} \\ \mu_1 : &\quad MSE_{SD} \neq MSE_{Benchmar} \end{aligned}$$

    The chance that this null hypothesis μ_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	$\lvert D_{trn} \rvert$	$\lvert D_{tst} \rvert$	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					
	100	70	50	40	30	20
Corr	0	0	0	0	0	0
HSIC-RFE	**30**	29	28	22	16	9
$\Delta\|\omega\|^2$-RFE	**30**	**30**	28	28	1	0
SD-L-RFE	**30**	**30**	**30**	**30**	**26**	**17**
SD-G-RFE	**30**	**30**	29	28	**26**	13

Method\\$|\mathcal{D}_{trn}|$	Additive Func				Interactive Func			
	200	100	70	50	200	100	70	50
Corr	15	8	5	3	4	3	2	1
HSIC-RFE	14	5	5	3	7	9	8	6
$\Delta\|\omega\|^2$-RFE	4	5	11	4	0	14	9	10
SD-L-RFE	**30**	27	21	**19**	**30**	**30**	29	**12**
SD-G-RFE	**30**	**28**	**23**	**19**	**30**	**30**	**30**	11

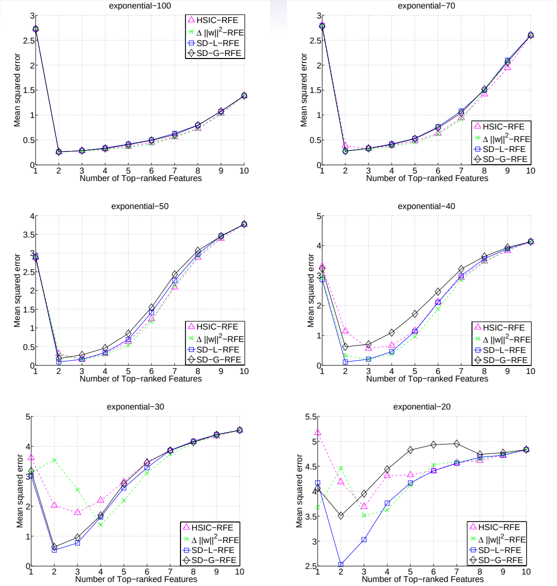


Figure: Average test MSE against top-ranked features over 30 realizations.

Real-World Problems

Table: Description of real-world data sets. C, κ and ϵ refer to SVR hyper-parameters C, κ, ϵ respectively.

Data sets	$\lvert\mathcal{D}_{trn}\rvert$	$\lvert\mathcal{D}_{tst}\rvert$	d	C	κ	ϵ
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}

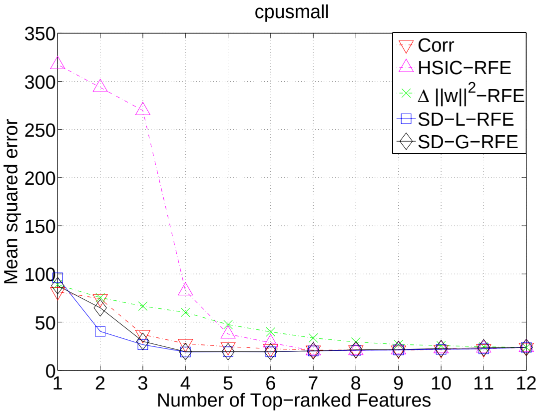

cpusmall
Mean squared error
Number of Top−ranked Features
Corr
HSIC−RFE
Δ ||w||^2−RFE
SD−L−RFE
SD−G−RFE
0
50
100
150
200
250
300
350
1
2
3
4
5
6
7
8
9
10
11
12

Table: t-test on data set cpusmall for 30 realizations

No.	SD-L-RFE	Corr		HSIC-RFE		$\Delta\|\omega\|^2$-RFE		SD-G-RFE	
	mean value	mean value	p-value	mean value	p-value	mean value	p-value	mean value	p-value
2	40.39	74.38	**0.00+**	293.6	**0.00+**	75.45	**0.00+**	64.81	**0.00+**
4	18.99	27.66	**0.00+**	82.44	**0.00+**	60.09	**0.00+**	19.33	0.55
6	19.20	22.33	**0.01+**	28.57	0.32	39.89	**0.00+**	19.22	0.97
8	20.66	21.09	0.49	20.49	0.78	29.36	**0.00+**	21.28	0.32
10	21.64	21.57	0.92	22.49	0.28	25.61	**0.00+**	22.52	0.24
12	23.78	23.78	1.00	23.78	1.00	23.78	1.00	23.78	1.00

Conclusions

- 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.