Spatial Decomposition for Large Scale SVMs

Philipp Thomann[1], Ingrid Blaschzyk[1], Mona Meister[2], Ingo Steinwart[1]

University of Stuttgart
Germany

[1]Institute for Stochastics and Applications
[2]BOSCH

Kernel Methods and their Problems for Large Scale Data

- Extensive theory, good performance for small and medium sized training sets
- Complexity for some domain-driven exponent $\alpha \in [1, 2]$:

$$O(n^2) \text{ in memory and } O(n^{1+\alpha}) \text{ in time}$$

e.g. if $n = 100\,000$ then kernel matrix needs 64GB of memory...

Old Idea: Random Chunks (RC)

Split data into random partitions of size k.

- Pros: kernel matrix has complexity $O(k \cdot n)$ and solver $O(k^\alpha \cdot n)$
- Cons: testing is $O(n \times n')$ for n' test samples and generalization error is poor and does not profit from more cells (see experiments).

Our Idea: Spatial Decompositions

SVMs are local: Partition input space and then train and test locally:

- Partition input space X into cells $\mathcal{A} := (A_j)_{j=1,\dots,m}$ and define local data set $D_j := \{(x, y) \in D : x \in A_j\}$.
- Train an individual support vector machine (SVM) on **each** cell.
- Test $x \in A_j$ by evaluating local decision function in x.

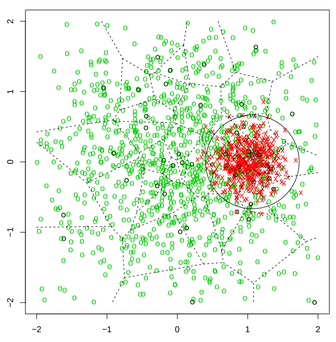
Now: memory complexity and solver complexity as for RC but also testing complexity $O(k \cdot n') \ll O(n \cdot n')$ and generalization error become better and better!

Algo	Kernel matrix (memory)	Solver (time)	Testing (time)	to improve error, increase
global	$O(d \cdot n^2)$	$O(n \cdot n^\alpha)$	$O(d \cdot n' \cdot n)$	n
RC	$O(d \cdot k \cdot n)$	$O(n \cdot k^\alpha)$	$O(d \cdot n' \cdot n)$	k
spatial	$O(d \cdot k \cdot n)$	$O(n \cdot k^\alpha)$	$O(d \cdot n' \cdot k)$	k and/or n

Table 1: Training set of size $d \times n$, test set of size n' and cells of mean size k.

Toy Example

Mixture of two Gaussians, one with labels 1, the other with labels -1:



Algorithm

Require: A training dataset D and a test set D'
Ensure: Test error

1: Find m centres (e.g. by farthest-first traversal).
2: **for all** $i = 1, \dots n$ **do**
3: Find nearest center j of x_i and add (x_i, y_i) to D_j.
4: **end for**
5: **for all** $j = 1, \dots, m$ **do**
6: Calculate SVM $f_{D_j,\lambda_j,\gamma_j}$.
7: **end for**
8: **for all** $i = 1, \dots n'$ **do**
9: Find nearest center j of x'_i and add (x'_i, y'_i) to D'_j.
10: **end for**
11: **for all** $j = 1, \dots, m$ **do**
12: Calculate test error $\mathcal{R}_{L,D'_j}(f_{D_j,\lambda_j,\gamma_j})$ on test cell D'_j.
13: **end for**
14: **return** global test error $\frac{1}{|D'|} \sum_{j=1}^m |D'_j| \cdot \mathcal{R}_{L,D'_j}(f_{D_j,\lambda_j,\gamma_j})$.

Experiments

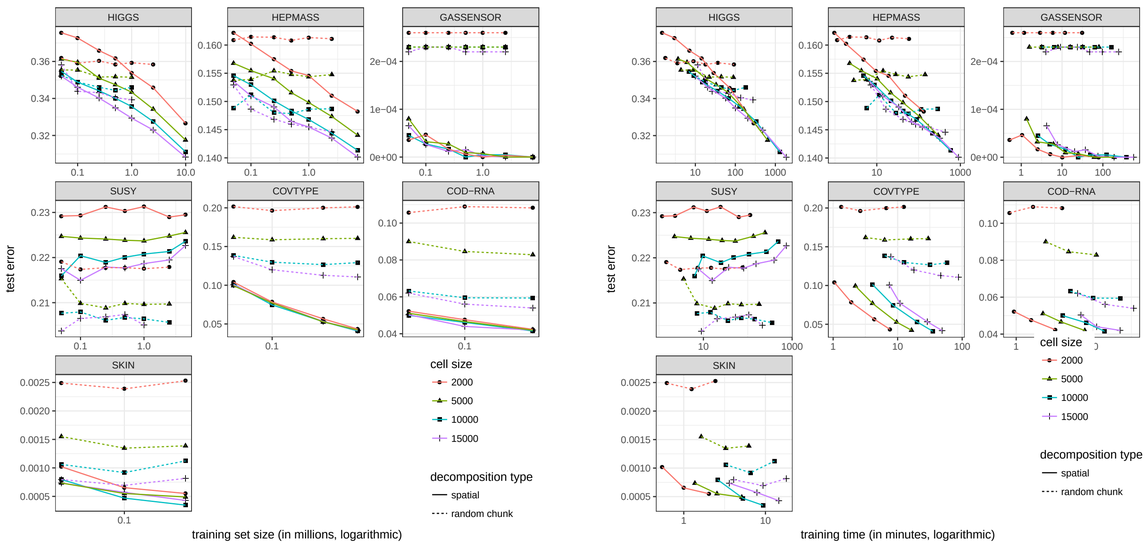


Figure 1: Training set size and training time. All times include 5-fold cross-validation on a 10×10-hyperparameter grid.

	size (mio.)	dim	2000	5000	10000	15000
			training time (in min.)			
HIGGS	10.0	28	308	679	1358	1992
HEPMASS	7.0	28	145	316	624	964
GASSENSOR	7.5	18	90	214	421	636
SUSY	4.5	18	121	261	513	779
COVTYPE	0.5	54	9	18	39	54
COD-RNA	0.2	9	4	9	16	25
SKIN	0.2	3	2	6	10	16

name	time (in sec.) used in phase						RAM
($k = 2000$)	part.	kernel calc.	solver	valid.	sel	test	
HIGGS	40	2887	11862	994	1030	322	31
HEPMASS	26	2022	4235	682	560	208	22
GASSENSOR	21	2354	729	633	541	166	21
SUSY	18	1296	4419	448	374	141	13
COVTYPE	3	138	206	44	47	13	3
COD-RNA	1	69	76	20	21	7	1
SKIN	1	53	21	18	11	3	1

Table 2: Detailed Times for Spatial Decompositions.

All experiments used `liquidSVM`, see `http://www.isa.uni-stuttgart.de/software/`.

The Math behind

- Data set $D := ((x_1, y_1), \dots, (x_n, y_n))$ drawn i.i.d. from a probability measure P.
- Learn an individual SVM on *each* cell by solving for a regularization parameter $\lambda_j > 0$ the optimization problem

$$f_{D_j,\lambda_j,\gamma_j} = \arg\min_{f \in H_{\gamma_j}} \lambda_j \|f\|^2_{H_{\gamma_j}} + \frac{1}{n} \sum_{x_i,y_i \in D_j} L(y_i, f(x_i)),$$

where
  - $L : Y \times \mathbb{R} \to [0, \infty)$ is a convex loss function,
  - H_{γ_j} is an RKHS over A_j with kernel parameter γ_j.
- Define the final decision function $f_{D,\boldsymbol{\lambda},\boldsymbol{\gamma}} : X \to \mathbb{R}$ by

$$f_{D,\boldsymbol{\lambda},\boldsymbol{\gamma}}(x) := \sum_{j=1}^m \mathbf{1}_{A_j}(x) f_{D_j,\lambda_j,\gamma_j}(x), \tag{1}$$

where $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_m)$ and $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_m)$.

Assumption on Partition

- Let $\mathcal{A}$ be such that $\mathring{A}_j \neq \emptyset$ for all $A_j \in \mathcal{A}$ and such that for $r > 0$ we have

$$r_{A_j} < r \leq 16 m^{-\frac{1}{d}}. \tag{2}$$

- **Toy Example**: Voronoi partition fulfils condition (2).

Margin Conditions

- Define $\eta(x) := P(y = 1|x)$, $x \in X$ and the sets

$$X_1 := \{ x \in X : \eta(x) > 1/2 \},$$
$$X_{-1} := \{ x \in X : \eta(x) < 1/2 \}.$$

- We call the function $\Delta_\eta : X \to [0, \infty)$, where

$$\Delta_\eta(x) := \begin{cases} \operatorname{dist}(x, X_1) & \text{if } x \in X_{-1}, \\ \operatorname{dist}(x, X_{-1}) & \text{if } x \in X_1, \\ 0 & \text{otherwise,} \end{cases}$$

distance to the decision boundary (DB), where $\operatorname{dist}(x, A) := \inf_{x' \in A} \|x - x'\|_2$.
- P has margin-noise exponent $\beta \in (0, \infty]$ if there exists a $c_{\text{MNE}} \geq 1$ such that for all $t > 0$

$$\int_{\{\Delta_\eta(x) < t\}} |2\eta(x) - 1| \, dP_X(x) \leq (c_{\text{MNE}} t)^\beta.$$

- P has (Tsybakov) noise exponent $q \in [0, \infty]$ if there exists a $c_{\text{NE}} > 0$ such that for all $t > 0$

$$P_X(\{|2\eta(x) - 1| < t\}) \leq (c_{\text{NE}} t)^q.$$

- **Toy Example:** $q = 1$ and $\beta = 2$.

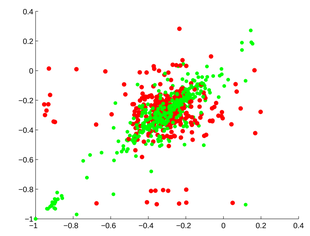


Figure 2: Almost no samples around the DB yield large β (bottom left); high concentration with malicious noise around the DB gives low β (center).

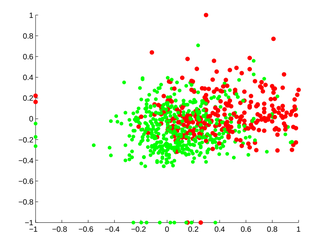


Figure 3: This 2-dim. projection of the diabetis data set has small noise exponent q. The margin-noise exponent β is not small since there is noise around DB.

Main Assumptions

- P has margin-noise exponent β and noise exponent q.
- Let assumption (2) on the partition hold. Let L be the hinge loss and let H_{γ_j} be an RKHS over A_j with Gaussian kernel parameter $\gamma_j \leq r$.

Finite Sample Bound

For all $p \in (0, 1)$, $n \geq 1$, $\tau \geq 1$ with $\tau \leq n$ the SVM in (1) satisfies

$$\mathcal{R}_{L,P}(f_{D,\lambda,\gamma}) - \mathcal{R}^*_{L,P} \leq C \left(\sum_{j=1}^m \lambda_j \gamma_j^{-d} + \gamma_{\max}^\beta + \left(\frac{\tau}{n}\right)^{\frac{q+1}{q+2}} + \left(\frac{r^{2p}}{n} \left(\sum_{j=1}^m \lambda_j^{-1} \gamma_j^{-\frac{d+2p}{p}} P_X(A_j) \right)^p \right)^{\frac{q+1}{q+2-p}} \right)$$

with probability P^n not less than $1 - 3e^{-\tau}$. The constant $C > 0$ depends only on d, β, p and q.

Learning Rate

- For a suitable choice of r_n, λ_n and γ_n we achieve for the local SVM the rate

$$n^{-\frac{\beta(q+1)}{\beta(q+2)+d(q+1)}+\xi},$$

where ξ can be chosen arbitrary small. **Toy Example**: Yields rate $n^{-\frac{2}{3+d}+\varepsilon}$.
- Rate coincides always with the fastest known rate which can be achieved by a global SVM.
- Rates can be obtained by a data-dependent parameter selection strategy without knowing the parameters.

References

[1] P. Thomann, I. Blaschzyk, M. Meister, and I. Steinwart.
Spatial decompositions for large scale svms.
In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, pages 1329–1337, 2017.