Outlier Removal for Enhancing Kernel-Based Classifier via the Discriminant Information

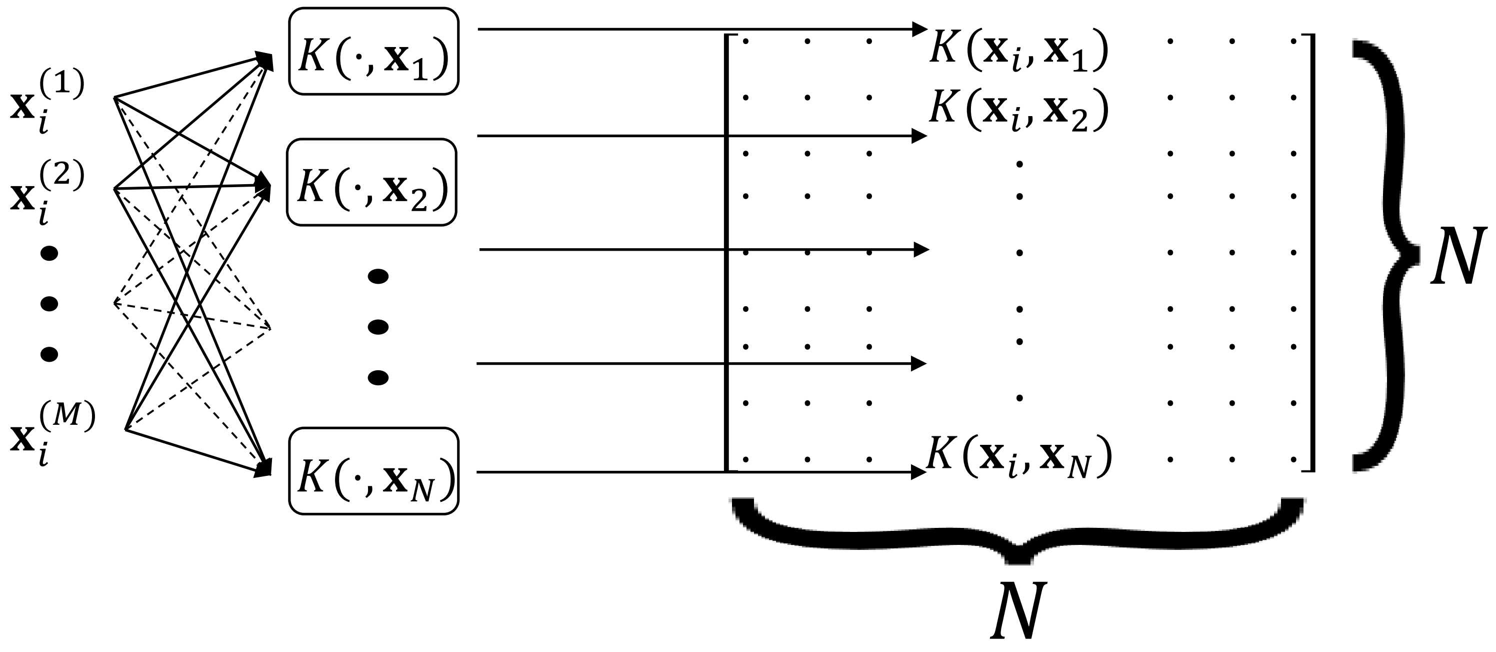
Thee Chanyaswad, Mert Al, and Sun-Yuan Kung

Department of Electrical Engineering

PRINCETON UNIVERSITY

Motivation

- Kernel methods have been successful techniques in pattern recognition for a variety of applications, e.g. speech, image, and medical diagnosis.
- However, kernel based learning has at least quadratic complexity in the number of training samples, which makes their use in large scale applications problematic.



Motivation

- There are two primary approaches for efficient, large-scale kernel based learning:
  - Kernel approximation (e.g. Nyström and Random Fourier features)
  - Data sample selection or outlying data removal.
- In order to be useful for scalable learning, these methods need to be computationally efficient.
- Our work follows the sample selection approach, and we use a filtering method to remove outlying samples.
- Our method requires only linear time to assign scores to all the data samples, even in a kernel induced feature space.

Discriminant Information (DI)

- Given a supervised training dataset $\{\mathbf{X} \in \mathcal{R}^{\boldsymbol{M} \times \boldsymbol{N}},\ \mathbf{y} \in \mathcal{R}^{\boldsymbol{N}}\}$ with N samples and M features, the discriminant information ψ is defined as

$$\psi = \mathrm{trace}\left((\bar{\mathbf{S}} + \rho \mathbf{I})^{-1} \mathbf{S}_B\right)$$

  where $\bar{\mathbf{S}}$ and $\mathbf{S}_B$ are the scatter matrix and between-class scatter matrix, respectively.

- We can obtain an equivalent expression for DI via the kernel trick,

$$\psi = \mathrm{trace}\left((\bar{\mathbf{K}}^2 + \rho \bar{\mathbf{K}})^{-1} \mathbf{K}_B\right)$$

  where $\bar{\mathbf{K}}$ is the centered kernel matrix and $\mathbf{K}_B$ is the kernel between-class scatter matrix.

Discriminant Information (DI)

- DI measures the separability of the data for classification:
  - Equals to zero when the class centers overlap (no separability).
  - Close to $L - 1$, where L is the number of data classes, when the samples are concentrated around their class centers (good separability).
- DI is indicative of a learner's classification ability, as demonstrated in earlier work.

1. S. Y. Kung, "Compressive privacy: From information /estimation theory to machine learning [lecture notes]," *IEEE Signal Processing Magazine*, vol. 34, no. 1, pp. 94–112, 2017.
2. Thee Chanyaswad, Mert Al, J. Morris Chang, and S. Y. Kung, "Differential mutual information forward search for multikernel discriminant-component selection with an application to privacy-preserving classification," in *Machine Learning for Signal Processing (MLSP), 2017 IEEE 27th International Workshop on*. 2017, IEEE.

Outlier Removal Discriminant Information (ORDI)

- Suppose the sample $(\mathbf{x}, y)$ has been removed from the training dataset. Let $\bar{\mathbf{S}}'$ and $\mathbf{S}_B'$ denote the scatter matrix and between-class scatter matrix obtained from the remaining data. We define the ORDI, $\partial\psi$ of the sample $(\mathbf{x}, y)$ as

$$\partial\psi = \mathrm{trace}\big((\bar{\mathbf{S}} + \rho\mathbf{I})^{-1}\mathbf{S}_B\big) - \mathrm{trace}\big((\bar{\mathbf{S}}' + \rho\mathbf{I})^{-1}\mathbf{S}_B'\big)$$

- We can similarly define ORDI with kernel matrices as

$$\partial\psi = \mathrm{trace}\big((\bar{\mathbf{K}}^2 + \rho\bar{\mathbf{K}})^{-1}\mathbf{K}_B\big) - \mathrm{trace}\left(\left({\bar{\mathbf{K}}'}^2 + \rho{\bar{\mathbf{K}}'}^2\right)^{-1}\mathbf{K}_B'\right)$$

- ORDI is expected to be small for outliers. Whereas it is expected to be large for samples that are easily separated from other classes.

Outlier Removal Discriminant Information (ORDI)

- Computing ORDI for a single sample can take $O(N^3)$ time and $O(N^2)$ memory.
- Need to find a criterion that is faster to compute, without compromising the predictive performance.

Bounding ORDI

- **Theorem:** Given a supervised training dataset $\{\mathbf{X} \in \mathcal{R}^{\boldsymbol{M}\times\boldsymbol{N}},\ \mathbf{y} \in \mathcal{R}^{\boldsymbol{N}}\}$ and a kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$, the Outlier Removal Discriminant Information of the sample $(\mathbf{x}, \mathrm{y})$ is bounded by

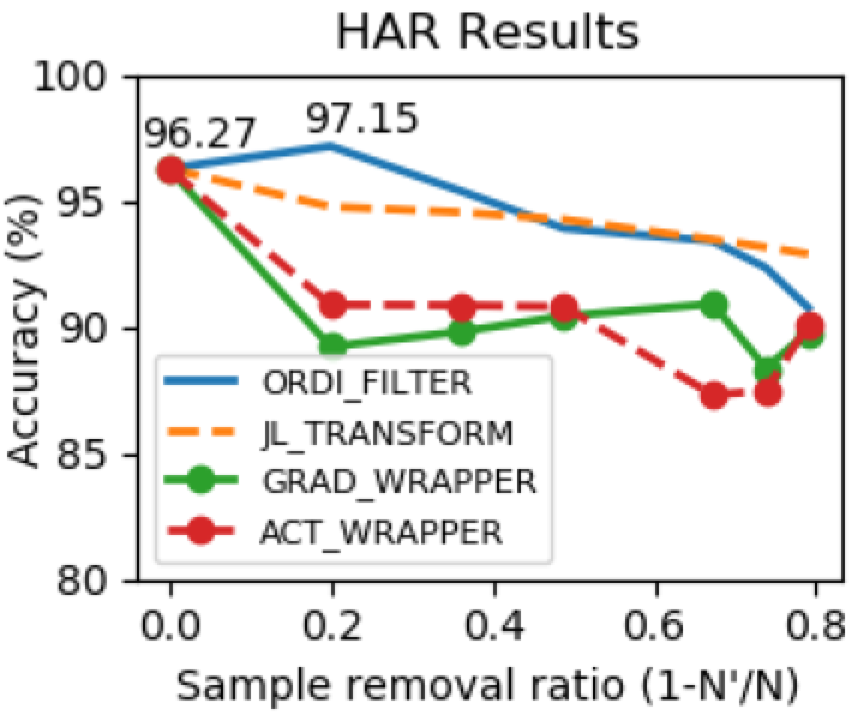
$$\partial\psi(\mathbf{x}, y) = \frac{\beta\kappa_{\mathbf{x}}}{\rho(\kappa_{\mathbf{x}} - \rho)} + \frac{H_{4,1/2}\left(\delta_{\mathbf{x},y} + \kappa_{\mathbf{x}}\right)}{\rho\left(N_y - 1\right)} + \frac{\kappa_{\mathbf{x}}\left(\delta_{\mathbf{x},y} + \kappa_{\mathbf{x}}\right)}{\rho(\kappa_{\mathbf{x}} - \rho)\left(N_y - 1\right)}$$

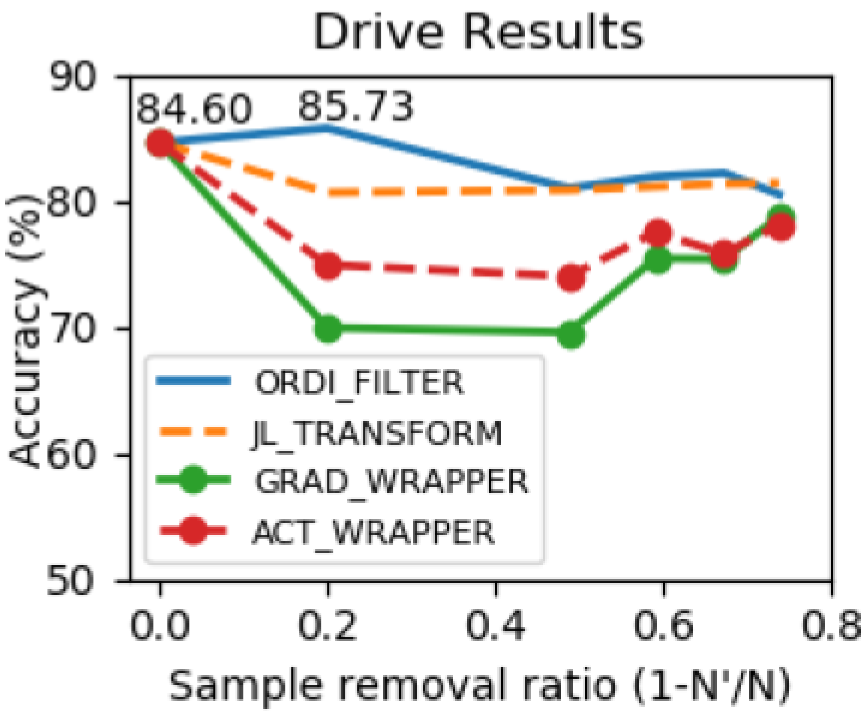
  where $\beta = \sum_{l=1}^{L} N_l k(\boldsymbol{\mu}_l, \boldsymbol{\mu}_l)$, $\kappa_{\mathbf{x}} = k(\mathbf{x}, \mathbf{x})$, N_l is the number of training samples in class l, $\boldsymbol{\mu}_l$ is the mean of the samples in class l, $H_{4,1/2}$ is the generalized harmonic number, and

$$\delta_{\mathbf{x},y} = N_y\left(k^2(\boldsymbol{\mu}_y, \boldsymbol{\mu}_y) - 4k(\boldsymbol{\mu}_y, \boldsymbol{\mu}_y)k(\mathbf{x}, \boldsymbol{\mu}_y) + 2\kappa_{\mathbf{x}}k(\boldsymbol{\mu}_y, \boldsymbol{\mu}_y) + 2k^2(\mathbf{x}, \boldsymbol{\mu}_y)\right)^{1/2}.$$

- Computing the upper bound on ORDI over the entire dataset requires only $O(N)$ time!

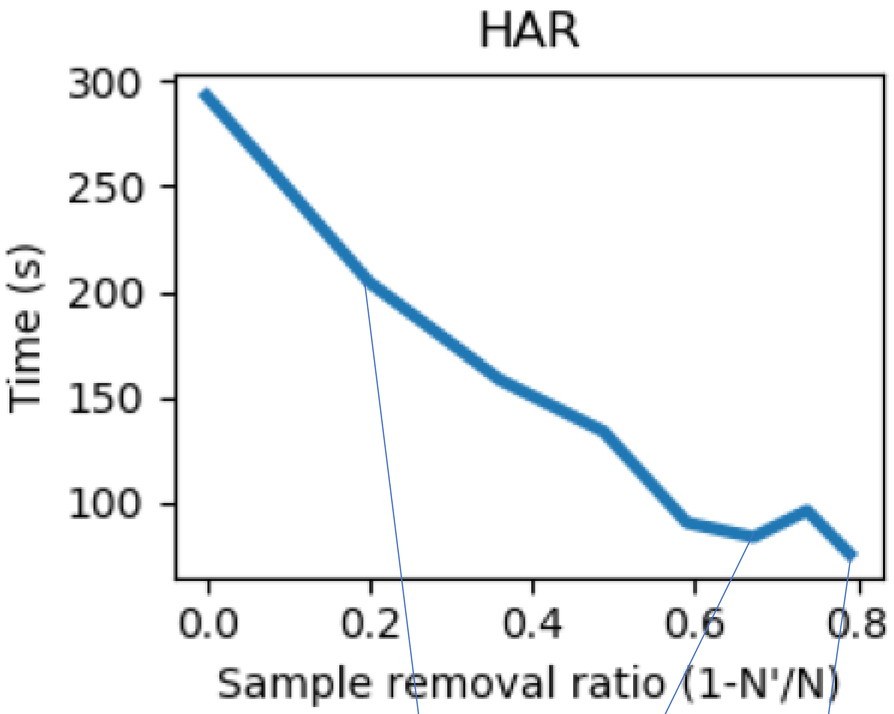
Experiments (Sample Ratios)

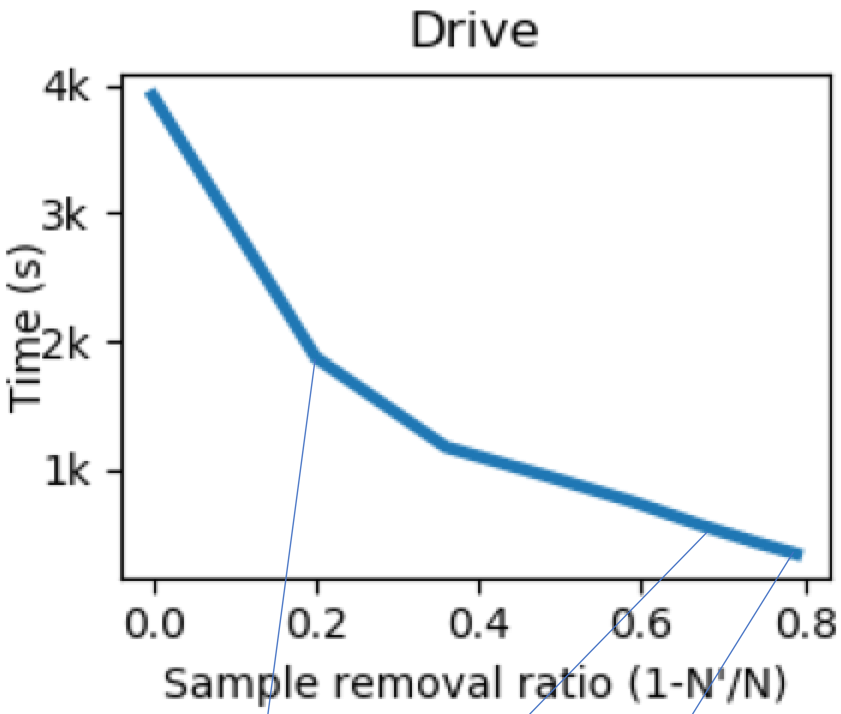




- We compare our filtering method based on the bound on ORDI with two wrapper methods and one filtering method for outlier sample removal.

Experiments (Training Times)





Time gains:	1.5x	3.5x	3.9x	2x	7x	9x
Accuracies:	+0.88%	-2,89%	-5.53%	+1.13%	-2.43%	-4.13%

Conclusion

- We proposed a filter approach for outlying data removal and sample selection in supervised learning, which only requires linear time to compute all the sample scores.
  - By removing 20% of the samples, we were able to exceed the performance of the original classifier on our two datasets, which leads to a win-win in terms of predictive performance and computational/memory cost.
  - By removing up to 80% of the samples, we were able to achieve very significant computational savings, by sacrificing relatively little accuracy.

References

1. S. Y. Kung, "Compressive privacy: From information /estimation theory to machine learning [lecture notes]," *IEEE Signal Processing Magazine*, vol. 34, no. 1, pp. 94–112, 2017.
2. Thee Chanyaswad, Mert Al, J. Morris Chang, and S. Y. Kung, "Differential mutual information forward search for multikernel discriminant-component selection with an application to privacy-preserving classification," in *Machine Learning for Signal Processing (MLSP), 2017 IEEE 27th International Workshop on*. 2017, IEEE.
3. S Y Kung, "A compressive privacy approach to generalized information bottleneck and privacy funnel problems," *Journal of the Franklin Institute*, 2017.
4. Thee Chanyaswad, J. Morris Chang, and Sun-Yuan Kung, "A compressive multi-kernel method for privacy-preserving machine learning," in *Neural Networks (IJCNN), 2017 International Joint Conference on*. 2017, pp. 4079–4086, IEEE.
5. Thee Chanyaswad, J. Morris Chang, Prateek Mittal, and SY Kung, "Discriminant-component eigenfaces for privacypreserving face recognition," in *Machine Learning for Signal Processing (MLSP), International Workshop on*. 2016, IEEE.