4-1-2019

Engineering Fast Multilevel Support Vector Machines

Ehsan Sadrfaridpour
Clemson University

Talayeh Razzaghi
Clemson University

Ilya Safro
Clemson University

ENGINEERING FAST MULTILEVEL SUPPORT VECTOR MACHINES

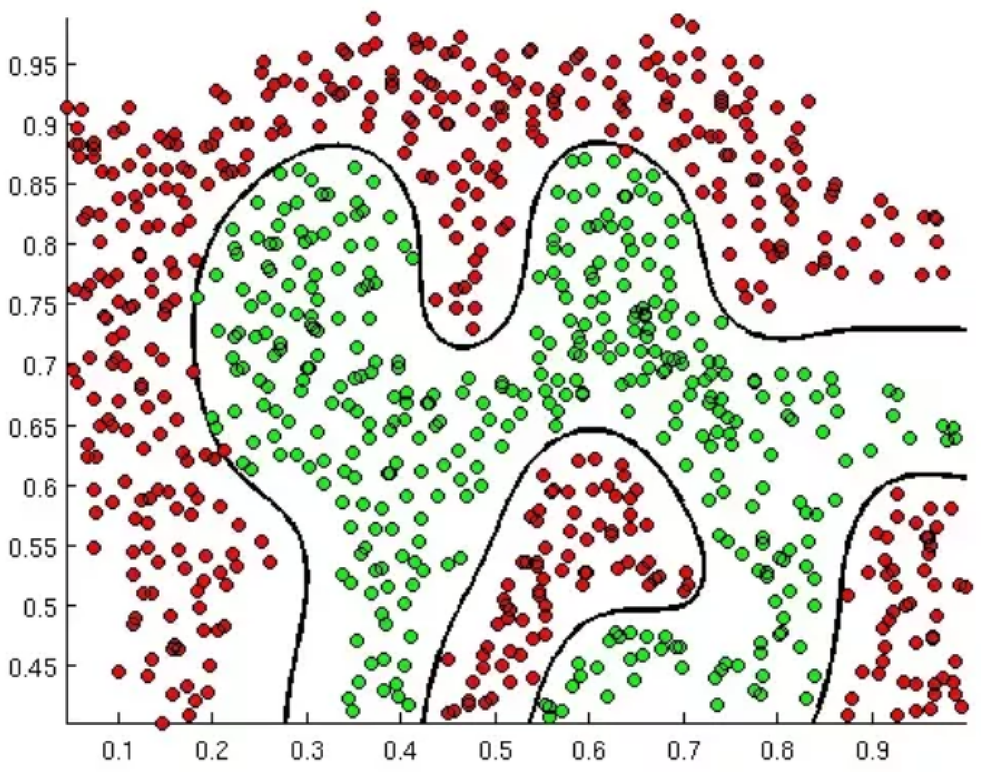
Ehsan Sadrfaridpour [1], TALAYEH RAZZAGHI [2], ILYA SAFRO [1]

1:SCHOOL OF COMPUTING, CLEMSON UNIVERSITY, 2: DEPARTMENT OF INDUSTRIAL ENGINEERING, NEW MEXICO STATE UNIVERSITY, LAS CRUCES, NM

CLEMSON UNIVERSITY

MOTIVATION

Classification Problem:



- Find a function to separate data points
- Accurately predict the unseen data
- Advance techniques are perform better for complex data points

MULTILEVEL SVM

- The problem:
  Solving large, hard, imbalanced data set **takes long time**
- Our solution:
  Very fast solver with high quality
- Source code: https://github.com/esadr/mlsvm

GENERAL (W)SVM

- Data points $\{(x_i, y_i)\}_{i=1}^{l}$
- Class labels (y_i) are $-1, +1$
- Weighted SVM: The WSVM is an extension of the SVM for imbalanced classes:

$$\min \ \frac{1}{2}\|w\|^2 + C^+ \sum_{i\in \mathbf{C}^+}^{n_+} \xi_i + C^- \sum_{j\in \mathbf{C}^-}^{n_-} \xi_j \quad (1a)$$

$$\text{s.t.} \ \ y_i(w^T\phi(x_i)+b) \geq 1-\xi_i \qquad i=1,\dots,l \quad (1b)$$

$$\xi_i \geq 0 \qquad i=1,\dots,l \quad (1c)$$

where the importance factors C^+, and C^- are associated with the positive, and negative classes respectively.

- The WSVM can be transformed into the Lagrangian dual and solved using the Kuhn-Tucker conditions.

REFERENCES

[1] Sadrfaridpour, E., Jeereddy, S., Kennedy, K., Luckow, A., Razzaghi, T., Safro, I. "Algebraic multigrid support vector machines", ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1

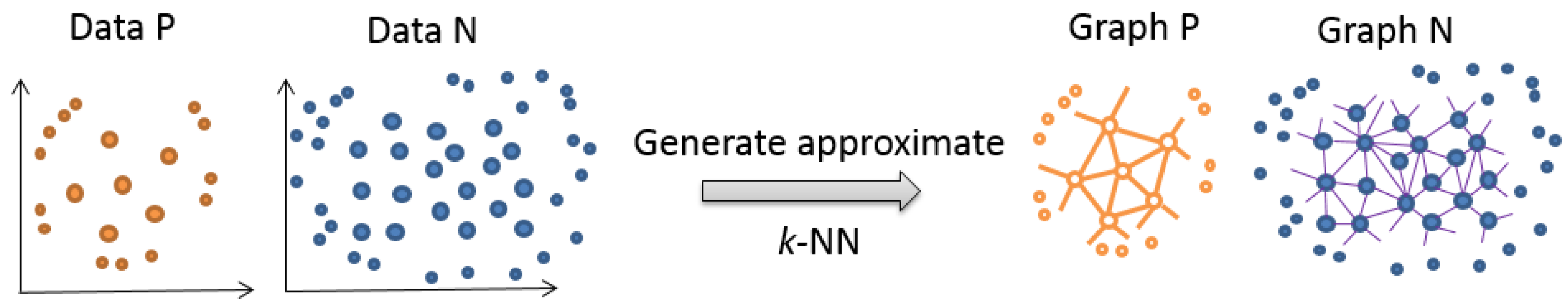
[2] Sadrfaridpour, E., Razzaghi, T., Safro, I. (2017). Engineering fast multilevel support vector machines. arXiv preprint arXiv:1707.07657.

[3] Razzaghi, T., Safro, I., Ewing, J., Sadrfaridpour, E., and Scott, J., D. (2019). "Predictive Models for Bariatric Surgery Risks with Imbalanced Medical Datasets." Annals of Operations Research, https://doi.org/10.1007/s10479-019-03156-8.

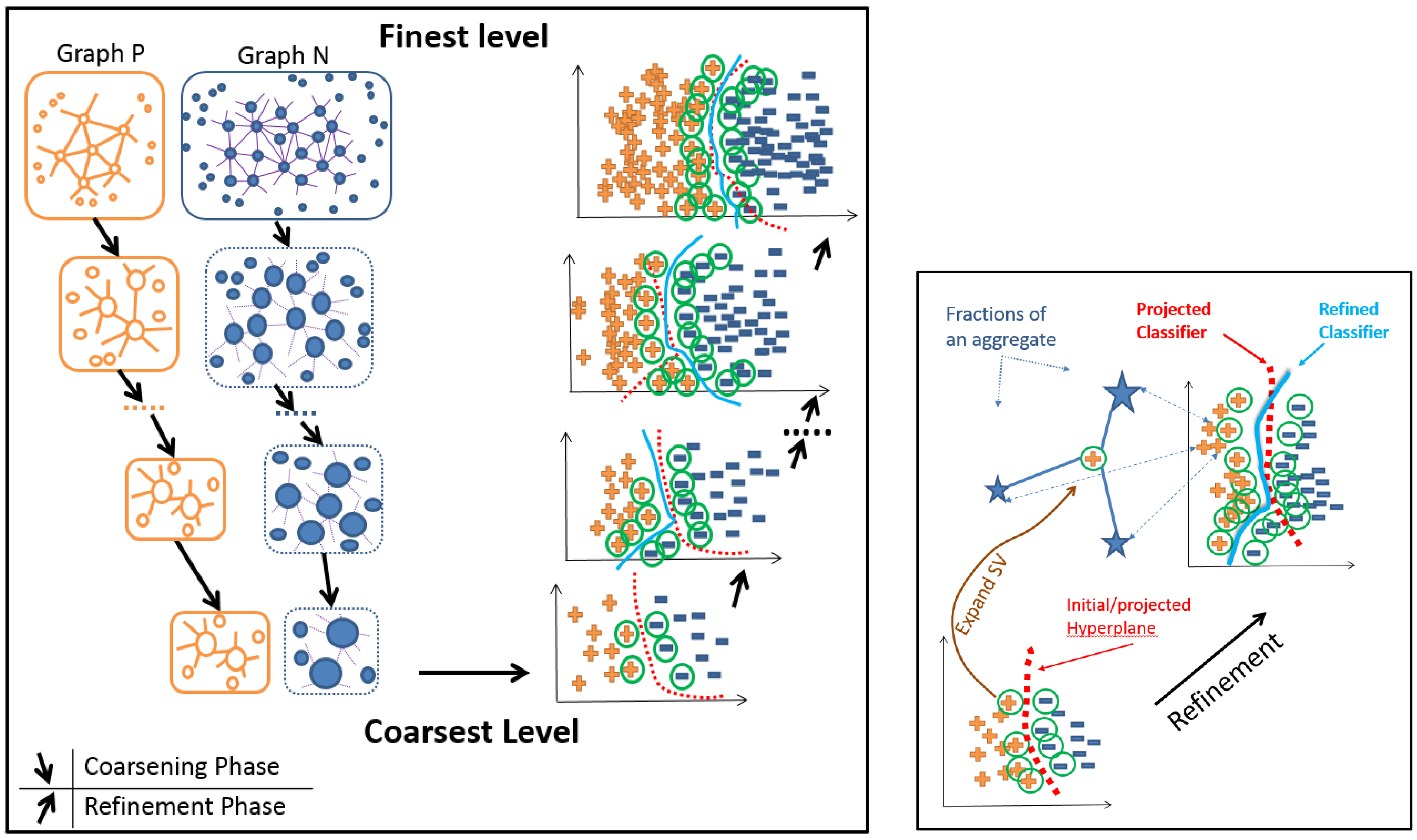
MULTILEVEL SUPPORT VECTOR MACHINES

We propose a new class of multilevel (W)SVM algorithms that is based on the **Algebraic Multigrid framework (AMG)**.

- **Coarse Problem**. The classification problem is reformulated as *(Weighted) Support Vector* classification problem for a non-uniform AMG coarsening, i.e., the C^+/C^- classes are separated.
- **AMG Coarsening**. We construct aggregates of fractions of data points using the Galerkin-like operator from AMG applied on the Laplacian of approximated k-NN graph.
- **Coarse Aggregates.** The construction of the set of aggregates is guided by the principle that each fine point should be "strongly coupled" to the chosen aggregates.



Creat k-NN graph from data points



(a) Overall Multilevel framework

(b) Adding SV's neighbors in a finer level

- **Refinement.** Add SVs and their neighbors in a finer level. In case of large training data, partition and choose closest pairwise partitions

RESULTS

- Our multilevel WSVM are evaluated on binary classification benchmark data sets.
- Performance measure (G-mean) and Time are reported for comparison between classical WSVM and MLSVM for small data sets and between LibLinear and MLSVM for large data sets
  r_{imb} is imbalance ratio, time in seconds
- **Main achievement: Fast computational time and improved quality on complex data sets**

Data set				Time (sec.)		G-mean	
Name	r_{imb}	f	*Size*	WSVM	MLSVM	WSVM	MLSVM
Advertisement	0.86	1558	3279	231	**213**	0.67	**0.91**
Buzz	0.80	77	140707	26026	**31**	0.89	**0.95**
Clean	0.85	166	6598	82	**5**	0.99	0.99
Cod-rna	0.67	8	59535	1857	**13**	**0.96**	0.94
Forest	0.98	54	581012	353210	**948**	**0.92**	0.88
Letter	0.96	16	20000	139	**30**	0.99	0.99
Nursery	0.67	8	12960	192	**2**	1.00	1.00
Ringnorm	0.50	20	7400	26	**2**	0.98	0.98
Twonorm	0.50	20	7400	28	**1**	0.98	0.98

Data set				Time (sec.)		G-mean	
Name	r_{imb}	f	*Size*	LibLinear	MLSVM	LibLinear	MLSVM
SUSY	0.54	18	5M	1300	**1116**	0.68	**0.74**
MNIST	0.90	784	4M	2626	**17581**	0.60	**0.84**
HIGGS	0.53	28	11M	4406	**3283**	0.54	**0.62**

ONGOING RESEARCH

- Develop the parallel version using MPI and OpenMP