

Algebraic multigrid support vector machines

Ehsan Sadrfaridpour¹, Sandeep Jeeredy², Ken Kennedy², Andre Luckow²
Talayah Razzaghi¹, Ilya Safro¹

1- Clemson University, School of Computing, Clemson SC, USA

2- Innovation Lab, BMW Group IT Research Center, Information Management Americas
Greenville SC, USA

November 18, 2016

Abstract

The support vector machine is a flexible optimization-based technique widely used for classification problems. In practice, its training part becomes computationally expensive on large-scale data sets because of such reasons as the complexity and number of iterations in parameter fitting methods, underlying optimization solvers, and nonlinearity of kernels. We introduce a fast multilevel framework for solving support vector machine models that is inspired by the algebraic multigrid. Significant improvement in the running has been achieved without any loss in the quality. The proposed technique is highly beneficial on imbalanced sets. We demonstrate computational results on publicly available and industrial data sets.

1 Introduction

Support vector machine (SVM) is one of the most well-known classification methods. The optimal classifier is achieved through solving a convex optimization model. Despite profound capability of SVM, the training process becomes highly time-consuming in two phases when the data is big, namely, training the classifier, and model selection. While training the classifier is a common phase in all SVMs, the model selection phase is usually applied on difficult data sets (e.g., when the data is noisy, imbalanced, and incomplete) in order to tune the parameters. On the one hand, SVM models are often much more flexible than other supervised classification methods. On the other hand, the flexibility comes with the price of finding the best model through tuning. Typically, the complexity of convex quadratic programming (QP) SVM algorithms is between $\mathcal{O}(n^2)$ to $\mathcal{O}(n^3)$ [11]. For example, the solver we compare our algorithm with, namely, LibSVM [4], which is one of the most popular QP solvers, scales between $\mathcal{O}(n_f n_s^2)$ to $\mathcal{O}(n_f n_s^3)$ subject to how effectively the cache is exploited in practice, where the numbers of features, and samples are denoted by n_f and n_s respectively. Clearly, this complexity is prohibitive for kernel based SVM models applied on practical big data without using parallelization and high-performance computing systems.

One of the major limitations of applying standard data mining classification algorithms is the imbalanced data, i.e., when the number of instances of one class is substantially greater than that in another class. In multi-class classification, the problem of imbalanced data is even bolder [18]. This might dramatically deteriorate the performance of algorithms. The SVM models are flexible enough to address the problem of imbalanced data. However, such models are usually computationally expensive. Since standard SVM algorithms often misclassify the data points of a small class, the cost-sensitive version of SVM, known as

weighted support vector machine (WSVM), has been developed. We are interested in developing a method that is scalable to very large data, and robust with respect to the imbalanced data.

In recent years, several strategies have been proposed to improve the performance of underlying QP solvers for big data. Efficient serial algorithms include decomposition techniques [22], shrinking and caching [13], and fast second order working set selection [9]. Another approach to accelerate the QP solvers is a chunking [13], in which the models is solved iteratively on the subsets of training data until the global optimum is achieved. A popular LibSVM solver [4] implements the sequential minimal optimization (SMO) algorithm. In the cases of easier data for which kernelized SVM is not required such approaches as LibLINEAR [8] exhibit good performance for linear SVMs using a coordinate descent algorithm. Another way to cope with the big data is through effective parallelization. In PSVM [32], the algorithm reduces memory use, and parallelizes data loading and computation in interior-point solver. Other works utilize many-core GPUs to accelerate SMO [23], and other architectures [31].

In this paper, we propose a novel method for efficient and effective solution of (W)SVM. In the heart of this method lies a multilevel algorithmic framework (MAF) inspired by the multiscale optimization strategies [2]. The main objective of MAF is to construct a hierarchy of problems (coarsening), each approximating the original problem but with fewer degrees of freedom. This is achieved by introducing a chain of successive restrictions of the problem domain into low-dimensional or small-size domains and solving the problem in them using local processing (uncoarsening). Typically, in computational optimization problems, the MAF combines solutions obtained by the local processing at different levels of coarseness into one global solution. Such frameworks have several key advantages that make them attractive for applying on large-scale data: they exhibit a linear complexity, and can be parallelized. Another advantage of the MAF is its heterogeneity, expressed in the ability to incorporate external appropriate optimization algorithms (as a refinement) in the framework at different levels. These frameworks were extremely successful in various practical machine learning and data mining tasks such as clustering [21, 16], segmentation [29], and dimensionality reduction [19].

Our contribution We introduce a novel multilevel framework for efficient and effective computation of (W)SVM classifiers. The algorithm is based on the algebraic multigrid (AMG) multilevel scheme [2]. We combine the AMG coarsening with the principles of: (a) coarse approximations of the support vectors, and (b) effective model selection parameter tuning through inheriting from the coarse scales. The framework is applied on both SVM and (W)SVM classifiers. To the best of our knowledge this is the first AMG-based algorithm for (W)SVM classification. The proposed method can be parallelized as any AMG algorithm, and its superiority is demonstrated on publicly available and industrial datasets of BMW. Our work extends and generalizes previous multilevel approaches such as [25, 24, 10] which results in a better running time and higher quality classifiers.

The major difference between typical computational optimization MAF, and the (W)SVM is the output of the model. In (W)SVM, the main output is the set of the support vectors which is usually much smaller than the total number of data points. We use this observation in our method by redefining the training set during the uncoarsening. In particular, we inherit the support vectors from the coarse scales, add their neighborhoods, and refine the support vectors at each fine scale. In other words, we improve the separating hyperplane throughout the hierarchy by gradual refinement of the support vectors until a global solution at the finest level is reached. In addition, we inherit the parameters of model selection and kernel from the coarse levels, and refine them throughout the uncoarsening.

2 Support Vector Machines

We briefly define the optimization problem underlying SVM models. Given n data points $\{x_i\}_{i=1}^n$ in \mathbb{R}^d , we define the corresponding labeled pairs (x_i, y_i) , where each x_i belongs to the class determined by the given label $y_i \in \{-1, 1\}$. Data points with positive labels are called “minority” class and are denoted by \mathbf{C}^+ , where $|\mathbf{C}^+| = n^+$. The rest of the points belongs to the “majority” class which is denoted by \mathbf{C}^- , where $|\mathbf{C}^-| = n^-$. Solving the following convex optimization problem by finding w , and b produces the hyperplane with maximum margin between \mathbf{C}^+ , and \mathbf{C}^-

$$\begin{aligned} & \text{minimize} && \frac{1}{2}\|w\|^2 + C \sum_{i=1}^n \xi_i && (1) \\ & \text{subject to} && y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \quad i = 1, \dots, n \\ & && \xi_i \geq 0, \quad i = 1, \dots, n. \end{aligned}$$

The mapping of data points to higher dimensional space is done by $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^p$ ($d \leq p$) to make two classes separable by a hyperplane. The term slack variables $\{\xi_i\}_{i=1}^n$ are used to penalize the misclassified points. The parameter $C > 0$ controls the magnitude of the penalization. The primal formulation is shown at (1) which is known as the *soft margin SVM* [30]. Solving the Lagrangian dual problem produces a reliable convergence which is faster than methods for primal formulation. The WSVM addresses imbalanced problems with assigning different weights to classes with parameters C^+ and C^- . The set of slack variables is split into two disjoint sets $\{\xi_i^+\}_{i=1}^{n^+}$, and $\{\xi_i^-\}_{i=1}^{n^-}$, respectively. In WSVM, the objective of (1) is changed to

$$\text{minimize} \quad \frac{1}{2}\|w\|^2 + C^+ \sum_{i=1}^{n^+} \xi_i^+ + C^- \sum_{j=1}^{n^-} \xi_j^- . \quad (2)$$

In all (W)SVM models, we use the Gaussian kernel $\exp(-\gamma\|x_i - x_j\|^2)$. Overall, in WSVM model, three parameters (C^+ , C^- , and γ) require tuning which is one of the main reasons of high complexity of these solvers. Typically, such parameter tuning techniques (such as the uniform design) apply sophisticated algorithms that iteratively run the solver many times to find the optimal parameters.

3 Algorithm

The goal of this paper is to introduce a framework to reduce the computational time of (W)SVM, while preserving and improving the quality of models. The proposed framework is inspired by the AMG solvers for computational optimization problems. It belongs to the family of multiscale hierarchical learning strategies with the following main phases: (a) coarsening; (b) coarsest scale learning; and (c) uncoarsening. In the coarsening process, the original problem is gradually restricted to smaller spaces by creating aggregates of fine data points and their fractions (an important feature of AMG), and turning them into data points at coarse levels. The main mechanism underlying the coarsening phase is the AMG which successfully helps to identify the interpolation operator for obtaining fine level solution from the coarse aggregates. When a hierarchy of coarse representations is created, and the number of coarse data points is sufficiently small, the coarsest scale learning is applied. In this stage, the (W)SVM problem is solved exactly on coarsest aggregates. In the uncoarsening phase, the solution obtained at the coarsest level (i.e., the support vectors and parameters) is gradually projected back to the finest level by interpolation and further local refinement of

support vectors and parameters. A critical difference between our approach and [25] is that in our approach the coarse support vectors are, in fact, not real data points prolonged from the finest level. Instead, they are aggregates of full fine-level data points and their fractions.

Framework initialization We initialize MAF with an undirected affinity graph $G = (V, E)$ generated from the training set of (W)SVM. Each data point i is associated with node $i \in V$ (same notation is used for points and nodes), and the set E is determined by the approximate k -nearest neighbor (k -NN) graph connections. We found a very little difference in the quality of the results if an exact k -NN graph is used while the running time for finding the approximate k -NN graph is significantly better. Throughout this paper, all approximate k -NN graphs are computed using FLANN library [20], where $k = 10$, and the distance is Euclidean. (We observed that increasing k does not improve the quality of the results.) This graph will serve as a structure for further coarsening.

In the multilevel graph frameworks [26], the edge weights represent the strength of connectivity between nodes in order to “simulate” the following interpolation scheme applied at the uncoarsening, in which strongly coupled nodes can interpolate solution to each other. For the classifier learning problems, this can be expressed as a similarity measure in the spirit of [16, 10], so we define a distance function between nodes (or corresponding data points) as an inverse of the Euclidean distance. Due to lack of space, we omit the results of experiments with other kernels which are currently being addressed in another paper as well as more advanced distance measure approaches such as [3, 5] that are often essential in multilevel methods.

In this paper, we work with binary classifiers (and one-vs-many multi-class classifiers) only but the approach is easily generalizable to multi-class classification. The coarsening is applied separately on both majority and minority classes, i.e., the C^+ points cannot be aggregated with points in C^- .

Coarsening Phase The main goal of the coarsening is to create a hierarchy of coarse representations of the original data manifold using the AMG coarsening for the approximated k -NN graph Laplacians. We denote the sequence of K next-coarser graphs by $\{G_i = (V_i, E_i)\}_{i=0}^K$, where $G_0 = G$ is the original graph that corresponds to the training set, and K is the number of levels in the hierarchy. For the completeness of the paper, we repeat the main steps of the AMG-based graph coarsening algorithm [28].

We describe a two-level process of obtaining the coarse graph $G_c = (V_c, E_c)$ and the corresponding coarse training set from the current fine level $G_f = (V_f, E_f)$ and its training set (e.g., the transition from level l to $l + 1$). The process is started with selecting seed nodes that will serve as centers of coarse level nodes, called aggregates. Coarse nodes will correspond to the coarse data points at level c . Structurally, each coarse aggregate can include one full seed f -level point, and possibly several other f -level points and their fractions. Intuitively, it is equivalent to grouping nodes in V_f into many small subsets allowing intersections, where each subset of nodes will correspond to a coarse point at level c . During the aggregation process, most coarse points will correspond to subsets of size greater than 1, so we introduce the notion of a volume $v_i \in \mathbb{R}_+$ for all $i \in V$ to reflect the importance of a point or its capacity that includes finest-level aggregated points and their fractions. We also introduce the edge weighting function $w : E_l \rightarrow \mathbb{R}_+$ for each graph G_l , $0 \leq l \leq K$, to reflect the strength of connectivity and similarity between nodes.

In Algorithm 1, we show the details of AMG coarsening. In the first step, we compute the future-volumes ϑ_i for all $i \in V_f$ to determine the order in which f -level nodes will be tested for declaring them as seeds (line 2), namely,

$$\vartheta_i = v_i + \sum_{j \in F} v_j \cdot \frac{w_{ji}}{\sum_{k \in V} w_{jk}}. \quad (3)$$

The future-volume ϑ_i is defined as a measure of how much an aggregate seeded by a data point i (or a node

in V_f) might potentially grow at the next level c .

We assume that in the finest level, all volumes are ones. We start with selecting a dominating set of seed nodes $C \subset V_f$ to initialize future coarse aggregates. Nodes that are not selected to C will belong to F such that $V_f = F \cup C$. Initially, the set F is set to be V_f , and $C = \emptyset$ since no seeds have been selected (line 1). After that, points with ϑ_i that is exceptionally larger than the average $\bar{\vartheta}$ are transferred to C as the most “representative” points (line 3). Then, all points in F are accessed in the decreasing order of ϑ_i updating C iteratively (lines 6-11), namely, if with the current C , and F , for point $i \in F$, $\sum_{j \in C} w_{ij} / \sum_{j \in V_f} w_{ij}$ is less than or equal to some threshold Q , i.e., the point is not strongly coupled to already selected points in C , then i is moved from F to C . The points with larger future-volumes usually have a better chance to be selected to C to serve as centers of future coarse points. Adding more seeds prevents too aggressive coarsening that can lead to “overcompressed” information at the coarse level and low quality classification model. However, it has been observed that in most AMG algorithms, $Q \geq 0.6$ is not required (however, this depends on the type and goals of aggregation). In our experiments $Q = 0.5$, and $\eta = 2$. Other similar values do not significantly change the results.

Algorithm 1 Selecting seeds for coarse nodes

```

1:  $C \leftarrow \emptyset, F \leftarrow V_f$ 
2: Calculate  $\forall i \in F \vartheta_i$ , and the average  $\bar{\vartheta}$ 
3:  $C \leftarrow$  nodes with  $\vartheta_i > \eta \cdot (\bar{\vartheta})$ 
4:  $F \leftarrow V_f \setminus C$ 
5: Recompute  $\vartheta_i \forall i \in F$ 
6: Sort  $F$  in descending order of  $\vartheta$ 
7: for  $i \in F$  do
8:   if  $\left( \frac{\sum_{j \in C} w_{ij}}{\sum_{j \in V_f} w_{ij}} \right) \leq Q$  then
9:     move  $i$  from  $F$  to  $C$ 
10:  end if
11: end for
12: return  $C$ 

```

When the set C is selected, we compute the AMG interpolation matrix $P \in \mathbb{R}^{|V_f| \times |C|}$ that is defined as

$$P_{ij} = \begin{cases} w_{ij} / \sum_{k \in N_i} w_{ik} & \text{if } i \in F, j \in N_i \\ 1 & \text{if } i \in C, j = I(i) \\ 0 & \text{otherwise} \end{cases}, \quad (4)$$

where $N_i = \{j \in C \mid ij \in E_f\}$ is the set of i th seed neighbors, and $I(i)$ denotes the index of a coarse point at level c that corresponds to a fine level aggregate around seed $i \in C$. Typically, in AMG methods, the number of non-zeros in each row is limited by the parameter called the interpolation order or caliber [2] (see discussion about R and Table 3). This parameter controls the complexity of a coarse-scale system (the number of non-zero elements in the matrix of coarse k -NN graph). It limits the number of fractions a fine point can be divided into (and thus attached to the coarse points). If a row in P contains too many non-zero elements then it is likely to increase the number of non-zeros in the coarse graph matrix. In multigrid methods, this number is usually controlled by different approaches that measure the strength of connectivity (or importance) between fine and coarse variables (see discussion and our implementation in [28]).

Using the matrix P , the aggregated data points and volumes for the coarse level are calculated. The edge between points $p = I(i)$ and $q = I(j)$ is assigned with weight $W_{pq}^{(coarse)} = \sum_{k \neq l} P_{ki} \cdot w_{kl} \cdot P_{lj}$. The volume for the aggregate $I(i)$ in the coarse graph is calculated by $\sum_j v_j P_{ji}$, i.e., the total volume of all points is preserved at all levels during the coarsening.

The stopping criteria for the coarsening depends on the available computational resources that can be used to learn the classifier at the coarsest level. In all our experiments, the coarsening stops when the size is less than a threshold (typically, 500 points) that ensures a fast performance of the LibSVM dual solver.

Note: ► One of the major advantages of the proposed coarsening scheme is the natural ability to deal with the imbalanced data. When the coarsening is performed on both classes simultaneously, and in a small class the number of points reaches an allowed minimum, this level is simply copied throughout the rest of levels required to coarsen the big class. Since the number of points at the coarsest level is small, this does not affect the overall complexity of the framework. ◀

Coarsest Level When both classes are small enough, the training reinforced by the parameter tuning is fast. We use the uniform design (UD) as a model selection technique to tune the parameters [12]. Another major advantage of the multilevel learning is the ability to inherit parameters C^+ , C^- , and γ during the uncoarsening. The tuned parameters are projected from the coarsest level back to next finer level, where they will be refined and projected up again. The coarsest level learning is shown in Algorithm 2.

Algorithm 2 Coarsest level i learning

- 1: **if** n^+ and n^- are sufficiently small for the coarsest level **then**
 - 2: Calculate the best $(S_i, C_i^+, C_i^-, \text{ and } \gamma_i)$ using UD, and (W)SVM solver on $|C_i^+|$ and $|C_i^-|$
 - 3: **Return** S_i (the set of coarsest support vectors), C_i^+ , C_i^- , and γ_i (learned parameters for level i)
 - 4: **end if**
-

Uncoarsening When the coarsest level is solved, we start to project the solution back gradually to the finest level. In contrast to the classical multilevel methods for computational optimization problems [2] in which each variable should be solved, the solution of (W)SVM consists of the set of support vectors whose size is typically much smaller than the number of data points. Thus, the main time-consuming ‘‘operation’’ of the uncoarsening is to project back and refine the set of coarse support vectors. This can be done very fast if we do not take into account all points at each level for the training. Instead, at each level, we define a new training set that includes only points from fine aggregates of the respective coarse level support vectors.

The $i + 1 \rightarrow i$ uncoarsening is presented in Algorithm 3. The set of support vectors S_{i+1} and parameters C_{i+1}^+ , C_{i+1}^- , and γ_{i+1} from level $i + 1$ are the inputs for level i . First, the new training data ($data_{\text{train}}$) is created by taking all level i points from the aggregates that correspond to the support vectors in S_{i+1} (lines 2-6). We denote by I^{-1} the reverse index function.

The parameter tuning using UD or other similar methods is a computationally expensive part of (W)SVM training which takes most of the time for large-scale data sets. Since it can be applied at the coarse levels of small size, we verify the size of a new $data_{\text{train}}$ (parameter Q_{dt}), and decide whether the UD is still applicable (line 7) or not. In case it can be applied, we run it around the parameters C_{i+1}^+ , C_{i+1}^- , and γ_{i+1} inherited from the coarse level $i + 1$ (lines 8-9). Otherwise, if the size of the training data is too large for the UD, we continue to inherit the parameters without adjusting them. Because in most problems, the number of support vectors is much smaller than the number of data points, even in very large data sets, we succeed to refine the parameters using UD at, approximately, 8-10 levels without any significant loss in the running

Table 1: Performance measures and running time (in seconds) for WSVM, and MLWSVM on publicly available data in [17].

Name	Datasets					WSVM					MLWSVM				
	r_{imb}	n_f	l	$ C^+ $	$ C^- $	ACC	SN	SP	κ	Time	ACC	SN	SP	κ	Time
Advertisement	0.86	1558	3279	459	2820	0.92	0.99	0.45	0.67	231	0.83	0.92	0.81	0.86	213
Buzz	0.80	77	140707	27775	112932	0.96	0.99	0.81	0.89	26026	0.88	0.97	0.86	0.91	233
Clean (Musk)	0.85	166	6598	1017	5581	1.00	1.00	0.98	0.99	82	0.97	0.97	0.97	0.97	7
Cod-RNA	0.67	8	59535	19845	39690	0.96	0.96	0.96	0.96	1857	0.94	0.97	0.92	0.95	102
Forest	0.98	54	581012	9493	571519	1.00	1.00	0.86	0.92	353210	0.88	0.92	0.88	0.90	479
Hypothyroid	0.94	21	3919	240	3679	0.99	1.00	0.75	0.86	3	0.98	0.83	0.99	0.91	3
Letter	0.96	16	20000	734	19266	1.00	1.00	0.97	0.99	139	0.98	1.00	0.97	0.99	12
Nursery	0.67	8	12960	4320	8640	1.00	1.00	1.00	1.00	192	1.00	1.00	1.00	1.00	2
Ringnorm	0.50	20	7400	3664	3736	0.98	0.99	0.98	0.98	26	0.98	0.98	0.98	0.98	2
Twonorm	0.50	20	7400	3703	3697	0.98	0.98	0.99	0.98	28	0.98	0.98	0.97	0.98	1

time. This gives us an effective and efficient practical parameter tuning technique that has been applied for several customer satisfaction classification problems in real-world large-scale data in recommender systems of BMW.

Algorithm 3 Uncoarsening from level $i + 1$ to i

```

1: Input:  $S_{i+1}, C_{i+1}^+, C_{i+1}^-, \gamma_{i+1}$ 
2:  $data_{train} \leftarrow \emptyset$ 
3: for all  $p \in S_{i+1}$  do
4:    $N_p \leftarrow$  all points in the aggregate  $I^{-1}(p)$ 
5:    $data_{train} \leftarrow data_{train} \cup N_p$ 
6: end for
7: if  $|data_{train}| < Q_{dt}$  then
8:    $C^O \leftarrow (C_{i+1}^+, C_{i+1}^-); \gamma^O \leftarrow \gamma_{i+1}$ 
9:    $(S_i, C_i^+, C_i^-, \text{and } \gamma_i) \leftarrow$  Run UD on (W)SVM using the initial center  $(C^O, \gamma^O)$ 
10: else
11:    $C_i^+ \leftarrow C_{i+1}^+$ 
12:    $C_i^- \leftarrow C_{i+1}^-$ 
13:    $\gamma_i \leftarrow \gamma_{i+1}$ 
14:    $S_i \leftarrow$  Apply (W)SVM on  $data_{train}$ 
15: end if
16: Return  $S_i, C_i^+, C_i^-, \text{and } \gamma_i$ 

```

The framework works in a similar way for both regular SVM and WSVM. The WSVM shows better performance for classification of the small class when the data is imbalanced.

4 Computational Results

The proposed framework is implemented in C++, and PETSc library which is the collection of data structures and methods for solving scientific computing problems [1]. PETSc provides a high-performance parallelization of algebraic structures that will be used in our future work that will be related to paralleliza-

Table 2: Evaluation of regular and multilevel WSVM for DS1 set of the COMPANY benchmark.

Class number	Size in DS1	Size in DS2	WSVM on DS1		MLWSVM on DS1		MLWSVM on DS2		
			ACC	κ	ACC	κ	ACC	κ	Time (in sec.)
Class 1	6867	204497	0.87	0.90	0.79	0.79	0.80	0.79	1123
Class 2	373	9892	0.99	0.36	0.90	0.69	0.63	0.69	200
Class 3	5350	91952	0.96	0.92	0.91	0.91	0.83	0.82	135
Class 4	278	9339	0.99	0.42	0.87	0.57	0.77	0.71	52
Class 5	2167	57478	0.93	0.62	0.63	0.69	0.62	0.66	53

Table 3: Quality of classifiers on publicly available data sets for different orders of interpolation.

Data set	κ						Time					
	R=1	R=2	R=4	R=6	R=8	R=10	R=1	R=2	R=4	R=6	R=8	R=10
Advertisement	0.86	0.80	0.84	0.84	0.86	0.82	219	205	220	205	213	268
Buzz	0.92	0.71	0.77	0.91	0.92	0.91	12	21	96	233	411	594
Clean (Musk)	0.96	0.96	0.95	0.97	0.96	0.97	6	7	7	7	8	8
Cod-RNA	0.94	0.95	0.95	0.95	0.95	0.94	48	140	84	59	146	150
Forest	0.63	0.51	0.59	0.90	0.89	0.85	84	68	168	479	1060	648
Hypothyroid	0.35	0.58	0.91	0.76	0.90	0.77	1	1	2	3	4	4
Letter	0.97	0.98	0.99	0.99	0.99	0.99	5	5	12	24	35	39
Nursery	1.00	1.00	1.00	1.00	1.00	1.00	2	3	3	3	4	5
Ringnorm	0.90	0.87	0.98	0.96	0.88	0.96	2	2	2	3	3	4
Twonorm	0.97	0.97	0.98	0.98	0.98	0.98	2	1	1	1	1	2

tion of the framework. Current implementation is not parallel. In general, based on the experience with with similar multilevel approaches [2], we anticipate the total complexity and performance of parallel version will be comparable to those of parallel AMG with small orders of interpolation. In our serial version, the linear complexity is comparable to serial AMG. The data structure we use are sparse matrices and vectors in the compressed row format. The rest of the data structures are STL of C++ 11. Small-scale (W)SVM models, that appear during the refinement, are solved using LibSVM 3.20 and the approximate k -NN graphs are constructed using FLANN.

To evaluate our algorithms, we use sensitivity (SN), specificity (SP), G-mean (κ), and accuracy (ACC), namely,

$$\text{SN} = \frac{TP}{TP + FN}, \quad \text{SP} = \frac{TN}{TN + FP}, \quad \kappa = \sqrt{\text{SP} \cdot \text{SN}} \quad (5)$$

and

$$\text{ACC} = \frac{TP + TN}{FP + TN + TP + FN}, \quad (6)$$

where TP , TN , FP , and FN are true positives, true negatives, false positives, and false negatives, respectively. We experimented with publicly available and real-workd industrial data of BMW. The publicly available data is available at the UCI collection [17]. The industrial data of recommendation systems is given in two data sets, namely, DS1, and DS2. They can also be available for limited research purposes. All computational results are averages over 20 similar executions with different random seeds, and randomly reordered data. The training-test split was 80%-20%.

In Table 1 (section “Datasets”), we present an information about the size of the data and its split into majority and minority classes. The notation r_{imb} , and n_f correspond to the imbalance factor, and the number of features, respectively. Performance measures of regular and multilevel WSVM are presented in sections WSVM, and MLWSVM of Table 1, respectively. Our main performance measure is κ since we are dealing with the imbalanced classification. We observed one significant improvement in the quality of κ in Advertisement data set. *In general, on these and several other data sets, no significant difference in the quality of κ between the proposed fast ML(W)SVM, and the full-time (W)SVM has been observed.*

The running time (in seconds) for both WSVM and MLWSVM is presented in columns “Time” in Table 1. The running time includes calculation of the approximated k -NN graphs and UD (model selection) for parameter tuning. *We demonstrate that the proposed fast AMG inspired framework justifies the idea of multilevel algorithms for (W)SVM, and clearly exhibits superior running time.*

Not surprisingly, it is much easier to analyze benchmarks like UCI machine learning dataset than the real-life industrial data which is very noisy, and contains missing values. In the BMW data, there are 5 labeled classes of plain text customer satisfaction surveys. First, the plain text is converted into normalized tf-idf form using the uni-, and bi-gram information which makes the number of features approximately 200.000 because of the extensive use of the domain-specific jargon. Then, we reduce the dimensionality of the data to 100 by applying SVD projections. *We note that we did not observe any change in the quality of the results for full, and reduced dimensional data except the increased running time for full dimensionality. While the multilevel (W)SVM framework running time is not fast but still realistic, the regular (W)SVM cannot be executed on such data at all without introducing significant changes such as high-performance parallelization or switching to linearized SVM version which significantly decreases the quality.*

The size of both DS1 and DS2 data sets is presented in columns 2-3, Table 2. Different classes (1-5) correspond to different major product problems addressed in the customer satisfaction surveys. For the evaluation of DS1 we focus only on the quality of the classifier because all running times are fast for this small dataset and mostly depend on the hardware, while for the DS2 set the running time is reported. While there is no loss in quality on both DS1, and DS2, the running time of MLWSVM on DS2 is substantially better than that of the regular WSVM which is measured in days, so it is comparable to the difference in running time of the Forest data set.

Does AMG help? One of the main reasons for developing a multilevel AMG-based SVM framework was an observation that for the real data of BMW, and experiments with complex healthcare data provided in [24], it is not enough to coarsen the data in the spirit of strict aggregation when data points are simply merged or eliminated based on some strong connectivity criteria such as in many clustering approaches [7, 10]. Applying other acceleration techniques such as an ensemble SVM learning [15, 6] also did not improve the quality of classifiers. We observed, that in many cases, the hyperplanes obtained at the coarse levels (i.e., without full uncoarsening) were substantially worse than the best known (but slowly computed) hyperplanes computed for the data sets that are known in the literature. Thus, we asked a question whether finding a better geometry of the data through more accurate AMG approximation of the spectral properties of the coarse approximated k -NN graphs can improve the quality of the classifier? We anticipated to have similar improvements to those obtained in segmentation [29], and clustering [16]. Unfortunately, because of several restrictions we cannot present full results of increasing interpolation order on the BMW data but we analyze them on public data sets.

In Table 3, we show the comparison of κ for data sets from [17] for different orders of interpolation (the number of non-zeros in rows of matrix P , see Eq. 4). It is easy to see that for the data sets Forest, and Hypothyroid the quality of classifier is improved for increased interpolation order R . Improvement of the quality comes with a price of increasing running time that is demonstrated in the “Time” section of Table 3.

Omitted observations (1) We are mostly interested in imbalanced problems, so we do not discuss the results of SVM and MLSVM, because their κ -quality is constantly worse than that of corresponding WSVM and MLWSVM. (2) We do not discuss a faster LibLINEAR solver [8] because of its significantly worse κ -quality. However, we note that if the data is not difficult enough, it can also be used as a part of the refinement instead of LibSVM. (3) We tested other solvers such as SVMlight [14]. No running time or quality improvements were observed.

5 Conclusions

We presented a new algorithmic framework for fast (W)SVM models. The framework belongs to the family of multiscale algorithms in which the problem is solved at multiple scales of coarseness, and gradually combined into one global solution of the original problem. We introduced the flexibility of the AMG coarsening and reinforced it with local learning of the support vectors and model selection parameters. This opens a number of interesting research directions to pursue. In particular, when the number of support vectors is indeed huge (which is not the case in many practical systems), we need to know how to combine multiple local hyperplanes into one global at the refinement stage that has to be applied locally for different clusters in the spirit of local refinement in other multiscale algorithms. Another major issue is related to effective inheritance scheme (such as bagging or ensemble SVM) of the model parameters for multiple hyperplanes. The implementation of our algorithms for ML(W)SVM is available at [27].

References

- [1] S Balay, S Abhyankar, M Adams, J Brown, P Brune, K Buschelman, V Eijkhout, W Gropp, D Kaushik, M Knepley, et al. *Petsc users manual revision 3.5*. Technical report, Technical report, Argonne National Laboratory (ANL), 2014.
- [2] A. Brandt and D. Ron. Chapter 1 : Multigrid solvers and multilevel optimization strategies. In J. Cong and J. R. Shinnerl, editors, *Multilevel Optimization and VLSICAD*. Kluwer, 2003.
- [3] J Brannick, M Brezina, S MacLachlan, T Manteuffel, S McCormick, and J Ruge. An energy-based amg coarsening strategy. *Numerical linear algebra with applications*, 13(2-3):133–148, 2006.
- [4] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.
- [5] Jie Chen and Ilya Safro. Algebraic distance on graphs. *SIAM J. Scientific Computing*, 33(6):3468–3490, 2011.
- [6] Marc Claesen, Frank De Smet, Johan AK Suykens, and Bart De Moor. Ensemblesvm: a library for ensemble learning using support vector machines. *Journal of Machine Learning Research*, 15(1):141–145, 2014.
- [7] Inderjit S Dhillon, Yuqiang Guan, and Brian Kulis. Weighted graph cuts without eigenvectors a multi-level approach. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(11):1944–1957, 2007.
- [8] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874, 2008.

- [9] Rong-En Fan, Pai-Hsuen Chen, and Chih-Jen Lin. Working set selection using second order information for training support vector machines. *The Journal of Machine Learning Research*, 6:1889–1918, 2005.
- [10] Haw-ren Fang, Sophia Sakellari, and Yousef Saad. Multilevel manifold learning with application to spectral clustering. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 419–428. ACM, 2010.
- [11] Hans P Graf, Eric Cosatto, Leon Bottou, Igor Dourdanovic, and Vladimir Vapnik. Parallel support vector machines: The cascade SVM. In *Advances in neural information processing systems*, pages 521–528, 2004.
- [12] C.M. Huang, Y.J. Lee, D.K.J. Lin, and S.Y. Huang. Model selection for support vector machines via uniform design. *Computational Statistics & Data Analysis*, 52(1):335–346, 2007.
- [13] Thorsten Joachims. Making large scale svm learning practical. Technical report, Universität Dortmund, 1999.
- [14] Thorsten Joachims. Svmlight: Support vector machine. *SVM-Light Support Vector Machine <http://svmlight.joachims.org/>*, University of Dortmund, 19(4), 1999.
- [15] Hyun-Chul Kim, Shaoning Pang, Hong-Mo Je, Daijin Kim, and Sung Yang Bang. Constructing support vector machine ensemble. *Pattern recognition*, 36(12):2757–2767, 2003.
- [16] Dan Kushnir, Meirav Galun, and Achi Brandt. Fast multiscale clustering and manifold identification. *Pattern Recognition*, 39(10):1876–1891, 2006.
- [17] M. Lichman. UCI machine learning repository, 2013.
- [18] Victoria López, Sara del Río, José Manuel Benítez, and Francisco Herrera. Cost-sensitive linguistic fuzzy rule based classification systems under the mapreduce framework for imbalanced big data. *Fuzzy Sets and Systems*, 258:5–38, 2015.
- [19] PietroGiorgio Lovaglio and Giorgio Vittadini. Multilevel dimensionality-reduction methods. *Statistical Methods & Applications*, 22(2):183–207, 2013.
- [20] Marius Muja and David G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *International Conference on Computer Vision Theory and Application VISSAPP'09*, pages 331–340. INSTICC Press, 2009.
- [21] Andreas Noack and Randolph Rotta. Multi-level algorithms for modularity clustering. In Jan Vahrenhold, editor, *Experimental Algorithms*, volume 5526 of *Lecture Notes in Computer Science*, pages 257–268. Springer Berlin Heidelberg, 2009.
- [22] Edgar Osuna, Robert Freund, and Federico Girosi. An improved training algorithm for support vector machines. In *Neural Networks for Signal Processing [1997] VII. Proceedings of the 1997 IEEE Workshop*, pages 276–285. IEEE, 1997.
- [23] John C Platt. Fast training of support vector machines using sequential minimal optimization. In *Advances in kernel methods*, pages 185–208. MIT press, 1999.

- [24] Talayeh Razzaghi, Oleg Roderick, Ilya Safro, and Nicholas Marko. Multilevel weighted support vector machine for classification on healthcare data with missing values. *PloS one*, 11(5):e0155119, 2016.
- [25] Talayeh Razzaghi and Ilya Safro. Scalable multilevel support vector machines. In *International Conference on Computational Science (ICCS)*, *Procedia Computer Science*, volume 51, pages 2683–2687. Elsevier, 2015.
- [26] Dorit Ron, Ilya Safro, and Achi Brandt. Relaxation-based coarsening and multiscale graph organization. *Multiscale Modeling & Simulation*, 9(1):407–423, 2011.
- [27] Ehsan Sadrfaridpour and Ilya Safro. AMG Support Vector Machines. <https://github.com/esadr/mlsvm>, 2016.
- [28] Ilya Safro, Dorit Ron, and Achi Brandt. Graph minimum linear arrangement by multilevel weighted edge contractions. *J. Algorithms*, 60(1):24–41, 2006.
- [29] Eitan Sharon, Meirav Galun, Dahlia Sharon, Ronen Basri, and Achi Brandt. Hierarchy and adaptivity in segmenting visual scenes. *Nature*, 442(7104):810–813, 2006.
- [30] Qiang Wu and Ding-Xuan Zhou. Svm soft margin classifiers: linear programming versus quadratic programming. *Neural computation*, 17(5):1160–1187, 2005.
- [31] Yang You, Haohuan Fu, Shuaiwen Leon Song, Amanda Randles, Darren Kerbyson, Andres Marquez, Guangwen Yang, and Adolfo Hoesie. Scaling support vector machines on modern hpc platforms. *Journal of Parallel and Distributed Computing*, 76:16–31, 2015.
- [32] Kaihua Zhu, Hao Wang, Hongjie Bai, Jian Li, Zhihuan Qiu, Hang Cui, and Edward Y Chang. Parallelizing support vector machines on distributed computers. In *Advances in Neural Information Processing Systems*, pages 257–264, 2008.