# A Dual Coordinate Descent Method for Large-scale Linear SVM

Kai-Wei Chang
Department of Computer Science
National Taiwan University

Joint work with C.-J. Hsieh, C.-J. Lin,
S. S. Keerthi, and S. Sundararajan

International Conference on Machine Learning, 2008

# Outline

- Introduction
- Dual Coordinate Descent
- Implementation Issue
- Comparisons
- Conclusions

# Outline

- Introduction
- Dual Coordinate Descent
- Implementation Issue
- Comparisons
- Conclusions

# Large-scale Linear Classifiers

Nonlinear SVM:

- SVM usually maps data into high dimensional space
- Hard to solve large data

Linear SVM:

- For applications like document classification
  Bag of words model (e.g., TF-IDF):
  large # of features
- Usually linear classifiers as good as kernelized ones
- Can solve larger problems than kernelized cases

# Large-scale Linear Classifiers (Cont'd)

Recently an active research topic

- [Keerthi and DeCoste, 2005, Lin et al., 2007]: Newton method
- [Joachims, 2006, Smola et al., 2008]: cutting plane
- [Shalev-Shwartz et al., 2007, Bottou, 2007]: stochastic gradient descent
- [Collins et al., 2008]: exponentiated gradient descent
- [Chang et al., 2008]: primal coordinate descent

# L1- and L2-SVM

- Training data $\{y_i, \mathbf{x}_i\}, \mathbf{x}_i \in R^n, i = 1, \ldots, l,\ y_i = \pm 1$
- L1-SVM:

$$\min_{\mathbf{w}} \ \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l} \max\left(0, 1 - y_i\mathbf{w}^T\mathbf{x}_i\right)$$

- L2-SVM:

$$\min_{\mathbf{w}} \ \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l} \left(\max(0, 1 - y_i\mathbf{w}^T\mathbf{x}_i)\right)^2$$

- But we solve the dual

# SVM Dual (Combining L1 and L2)

- From primal dual relationship

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & f(\boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\alpha}^T \bar{Q} \boldsymbol{\alpha} - \mathbf{e}^T \boldsymbol{\alpha} \\ \text{subject to} \quad & 0 \le \alpha_i \le U, \forall i, \end{aligned}$$

- $\bar{Q} = Q + D$
- $Q_{ij} = y_i y_j \mathbf{x}_i^T \mathbf{x}_j$
- $D$: diagonal matrix; $\mathbf{e}$: vector of all ones
- L1-SVM: $U = C$ and $D_{ii} = 0, \ \forall i$
- L2-SVM, $U = \infty$ and $D_{ii} = 1/(2C), \ \forall i$.

# Outline

# Dual Coordinate Descent

- Very simple: minimizing one variable at a time
- While $\boldsymbol{\alpha}$ not optimal

  For $i = 1, \ldots, l$

$$\min_{\alpha_i} f(\cdots, \alpha_i, \cdots)$$

- A classic optimization technique
- Traced back to [Hildreth, 1957]

  if constraints are not considered
- Studied by several SVM papers

# Dual Coordinate Descent (Cont'd)

- [Mangasarian and Musicant, 1999]:
  But didn't focus on linear SVM for large number of features
- Recently, [Bordes et al., 2007]
  For multi-class with kernels; didn't focus on linear SVM
- Others (e.g. [Crammer and Singer, 2003])
- We show a good coordinate descent implementation is very efficient for large linear SVM

# The Procedure

- Given current $\boldsymbol{\alpha}$. Let $\mathbf{e}_i = [0, \ldots, 0, 1, 0, \ldots, 0]^T$.

$$\min_{d} \; f(\boldsymbol{\alpha} + d\mathbf{e}_i) = \frac{1}{2}\bar{Q}_{ii} d^2 + \nabla_i f(\boldsymbol{\alpha}) d + \text{constant}$$

- Without constraints

$$\text{optimal } d = -\frac{\nabla_i f(\boldsymbol{\alpha})}{\bar{Q}_{ii}}$$

- Now $0 \leq \alpha_i + d \leq U$

$$\alpha_i \leftarrow \min\left(\max\left(\alpha_i - \frac{\nabla_i f(\boldsymbol{\alpha})}{\bar{Q}_{ii}}, 0\right), U\right)$$

# The Procedure (Cont'd)

$$\nabla_i f(\boldsymbol{\alpha}) = (\bar{Q}\boldsymbol{\alpha})_i - 1 = \sum_{j=1}^{l} Q_{ij}\alpha_j - 1 + D_{ii}\alpha_i$$

$$= \sum_{j=1}^{l} y_i y_j \mathbf{x}_i^T \mathbf{x}_j \alpha_j - 1 + D_{ii}\alpha_i$$

- Directly calculate gradients costs $O(ln)$
  $l$:# data, $n$: # features
- For linear SVM, define

$$\mathbf{w} = \sum_{j=1}^{l} y_j \alpha_j \mathbf{x}_j,$$

- Easy gradient calculation: costs $O(n)$

$$\nabla_i f(\boldsymbol{\alpha}) = y_i \mathbf{w}^T \mathbf{x}_i - 1 + D_{ii}\alpha_i$$

# The Procedure (Cont'd)

- All we need is to maintain $\mathbf{w}$

$$\mathbf{w} = \sum_{j=1}^{l} y_j \alpha_j \mathbf{x}_j,$$

- If

$$\bar{\alpha}_i : \text{ old } ; \qquad \alpha_i : \text{ new}$$

then

$$\mathbf{w} \leftarrow \mathbf{w} + (\alpha_i - \bar{\alpha}_i) y_i \mathbf{x}_i.$$

Also costs $O(n)$

- Sparse: $O(\#\text{nz})$ reduce to $O(\#\text{nz}/l)$

# Algorithm

- Given initial $\boldsymbol{\alpha}$ and the corresponding $\mathbf{w} = \sum_i y_i \alpha_i \mathbf{x}_i$.
- While $\boldsymbol{\alpha}$ is not optimal (Outer iteration)

  For $i = 1, \ldots, l$ (Inner iteration)

  (a) $\bar{\alpha}_i \leftarrow \alpha_i$

  (b) $G = y_i \mathbf{w}^T \mathbf{x}_i - 1 + D_{ii}\alpha_i$

  (c) If $\alpha_i$ can be changed

  $$\alpha_i \leftarrow \min(\max(\alpha_i - G/\bar{Q}_{ii}, 0), U)$$

  $$\mathbf{w} \leftarrow \mathbf{w} + (\alpha_i - \bar{\alpha}_i) y_i \mathbf{x}_i$$

# Analysis

- Convergence; extending results in [Luo and Tseng, 1992]

$$f(\boldsymbol{\alpha}^{k+1}) - f(\boldsymbol{\alpha}^*) \leq \mu(f(\boldsymbol{\alpha}^k) - f(\boldsymbol{\alpha}^*)), \forall k \geq k_0.$$

  $\boldsymbol{\alpha}^*$: optimal solution

- A careful implementation greatly improves the speed

# Outline

# Shrinking: Much Easier than Nonlinear

- Remove $\alpha_i$ if it is likely to be bounded until the end
  Smaller optimization problem
- Check stopping condition of the whole problem
  $\Rightarrow$ Need $\nabla f(\boldsymbol{\alpha})$
- Non-linear SVM: $O(l^2 n)$ to reconstruct gradient
- Linear: $\nabla f(\boldsymbol{\alpha})$ reconstructed by $\mathbf{w}$ $\Rightarrow$ only $O(ln)$

$$\mathbf{w} = \sum\nolimits_{\text{shrunken}} y_i \alpha_i \mathbf{x}_i + \sum\nolimits_{\text{other}} y_i \alpha_i \mathbf{x}_i$$

- Due to sequential updating, $O(ln)$ are not needed (details not shown)

# Order of Sub-problems

- Order of sub-problems being minimized

  $$\alpha_1 \rightarrow \alpha_2 \rightarrow \cdots \rightarrow \alpha_l$$

  Can use any random order at each outer iteration

  $$\alpha_{\pi(1)} \rightarrow \alpha_{\pi(2)} \rightarrow \cdots \rightarrow \alpha_{\pi(l)}$$

  Very effective in practice
- Online Setting: pick an $\alpha_i$ to update at once

  Related to

  [Collins et al., 2008, Crammer and Singer, 2003, Shalev-Shwartz et al., 2007]

# Outline

# Comparisons (Latest Version Used)

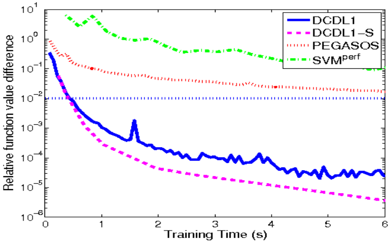
L1-SVM

- DCDL1: Dual coordinate descent (DCDL1-S: with shrinking)
- Pegasos [Shalev-Shwartz et al., 2007]: stochastic gradient descent
- SVM$^{\text{perf}}$ [Joachims, 2006]: cutting plane

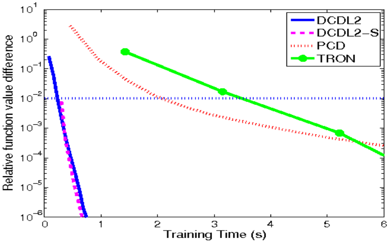
L2-SVM

- DCDL2: Dual coordinate descent (DCDL2-S: with shrinking)
- PCD [Chang et al., 2008]: Primal coordinate descent
- TRON [Lin et al., 2007]: Newton method

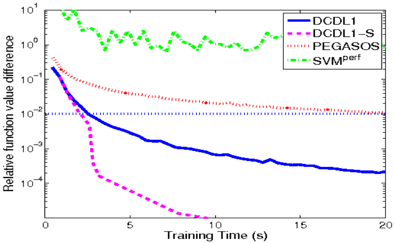
# Objective values (Time in Seconds)



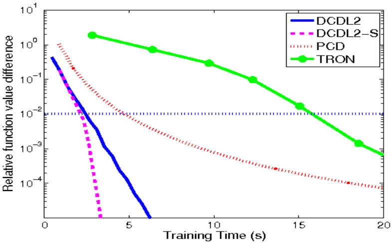

L1-SVM: news20




L2-SVM: news20




L1-SVM: rcv1



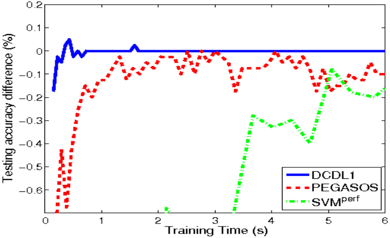

L2-SVM: rcv1

# Objective values (Time in Seconds)

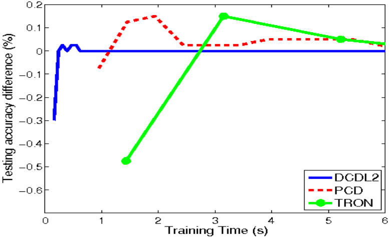
- Time for a solver to reduce the primal objective value to within 1% of the optimal value

| Data set | Linear L1-SVM | | | Linear L2-SVM | | |
|---|---|---|---|---|---|---|
| | DCDL1 | Pegasos | $\text{SVM}^{\text{perf}}$ | DCDL2 | PCD | TRON |
| astro-physic | 0.2 | 2.8 | 2.6 | 0.2 | 0.5 | 1.2 |
| real-sim | 0.2 | 2.4 | 2.4 | 0.1 | 0.2 | 0.9 |
| news20 | 0.5 | 10.3 | 20.0 | 0.2 | 2.4 | 5.2 |
| yahoo-japan | 1.1 | 12.7 | 69.4 | 1.0 | 2.9 | 38.2 |
| rcv1 | 2.6 | 21.9 | 72.0 | 2.7 | 5.1 | 18.6 |
| yahoo-korea | 8.3 | 79.7 | 656.8 | 7.1 | 18.4 | 286.1 |

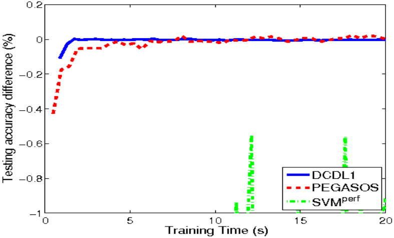
# Testing Accuracy (Time in Seconds)



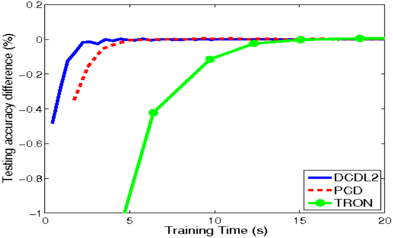

L1-SVM: news20




L2-SVM: news20




L1-SVM: rcv1




L2-SVM: rcv1

# Outline

- Introduction
- Dual Coordinate Descent
- Implementation Issue
- Comparisons
- Conclusions

# Conclusions

- Dual coordinate descents very effective if # data, # features large

  Useful for document classification
- Half million data in a few seconds
- Too good to be true? Any limitation?
- Less effective if

  # features small: should solve primal; or

  large penalty parameter $C$ (generally not needed for document data)

# Conclusions (Cont'd)

- Proposed methods included in the package LIBLINEAR

  `http://www.csie.ntu.edu.tw/~cjlin/liblinear`
- All sources used for experiments are available at the same page