Learning Sparse SVM for Feature Selection on Very High Dimensional Datasets

Mingkui Tan †, Li Wang†‡, Ivor W. Tsang†

† School of Computer Engineering, Nanyang Technological University, Singapore
‡ Department of Mathematics, University of California, San Diego, USA

ICML, June 24, 2010
Outline

1 Introduction to Large-Margin Based Feature Selection
   • Motivation of Feature Selection
   • Non-monotonic Feature Selection
   • $l_p$ Regularization
   • SVM-RFE
Outline

1 Introduction to Large-Margin Based Feature Selection
   - Motivation of Feature Selection
   - Non-monotonic Feature Selection
   - $l_p$ Regularization
   - SVM-RFE

2 Our Contributions
Outline

1. Introduction to Large-Margin Based Feature Selection
   - Motivation of Feature Selection
   - Non-monotonic Feature Selection
   - $l_p$ Regularization
   - SVM-RFE

2. Our Contributions

3. Feature Generating Machine (FGM)
   - New Sparse SVM Model
   - Convex Relaxations
   - Feature Generating Machine
   - Convergence and Complexity Analysis
Outline

1. Introduction to Large-Margin Based Feature Selection
   - Motivation of Feature Selection
   - Non-monotonic Feature Selection
   - $l_p$ Regularization
   - SVM-RFE

2. Our Contributions

3. Feature Generating Machine (FGM)
   - New Sparse SVM Model
   - Convex Relaxations
   - Feature Generating Machine
   - Convergence and Complexity Analysis

4. Experimental Results
Outline

1 Introduction to Large-Margin Based Feature Selection
   - Motivation of Feature Selection
   - Non-monotonic Feature Selection
   - $l_p$ Regularization
   - SVM-RFE

2 Our Contributions

3 Feature Generating Machine (FGM)
   - New Sparse SVM Model
   - Convex Relaxations
   - Feature Generating Machine
   - Convergence and Complexity Analysis

4 Experimental Results

5 Conclusions
Motivation of Feature Selection

- Reduce the "curse of dimensionality" problem.
- Remove the noninformative features and improve the generalization performance.
- Lead to simplified decision rule for faster predictions.
- Identify a small subset of features to better interpret the results.
Non-monotonic Feature Selection

- "Non-monotonic" feature selection means one needs to find the most informative feature subset combinations.
- "Monotonic" feature selection:
  If one informative feature is wrongly removed from a subset $S$, it will not be in its nested subsets [Xu et al.(2009)].
- Filter methods are monotonic methods.
- [Xu et al.(2009)] propose a "non-monotonic" MKL (NMMKL) for "non-monotonic" feature selection.
  - It learns the best feature combinations by solving a multiple linear kernel learning problem.
  - It preserves the "non-monotonic" property.
  - It is expensive in computation.
Given \( \{x_i, y_i\}_{i=1}^{n}, x_i \in \mathbb{R}^m \), \( l_p \) regularization minimizes the structural risk functional:

\[
\min_w \Omega(\|w\|_p) + C \sum_{i=1}^{n} \ell(-y_i w' x_i),
\]

(1.1)

Here \( \|w\|_p \) is the sparse regularizer and \( 0 \leq p \leq 1 \).

- When \( p = 0 \), it gives the \( l_0 \)-norm regularization.
  - Nonconvex and NP-hard.
- Convex \( l_0 \)-norm relaxations:
  - \( \text{AROM}[\text{Weston et al.}(2003)] \); \( \text{QCQP-SSVM} \) & \( \text{SDP-SSVM} \)
    - \( \text{AROM}[\text{Chan et al.}(2007)] \), ....
  - Too relax for \( l_0 \)-norm.
  - Computationally expensive.
When $p = 1$, it solves a linear programming problem (LPSVM) [Bradley & Mangasarian(1998), Fung & Mangasarian(2004)].

$$
\min_w \Omega(\|w\|_1) + C \sum_{i=1}^n \ell(-y_i w^T x_i) \quad (1.2)
$$

- Convex and can be easily solved (see in [Yuan et al]).
- Hard to control the sparsity (i.e., hard to choose $C$).
- Not so good for non-sparse problems (i.e., all features may contribute to the classification).
SVM-RFE

Based on the hyperplane $y = w'x$, SVM-RFE [Guyon et al.(2002)] recursively eliminates chunks of features with the least weights $w_i^2$.

- **Advantages:**
  - Shows good performance on small sample size problems.
  - Easy to be implemented.

- **Disadvantages:**
  - Hard to control the chunk size.
  - Not suitable for high dimensional dense problems (It takes $O(nm^2)$ time complexity when chunk size=1).
  - Monotonic and only local optimal.

SVM-RFE is used as the baseline method in our experiments.
Propose a new sparse model and then transform it into a QCQP problem with exponential constraints via a mild convex relaxation.

Propose to solve the relaxed problem by using cutting plane algorithm which incorporates the MKL learning.

The proposed algorithm, namely feature generating machine (FGM), can globally converge within a small number of iterations.

FGM is “non-monotonic”.

FGM scales linearly on both dimensions and instances.
New Sparse $l_2$-SVM Model

Introduce a “0-1” vector $d$ into the standard $l_2$-SVM to control the features. “1” denotes the feature being selected and “0” denotes not. Suppose $B$ features are selected, the scheme is as follows:

\[
\begin{align*}
\text{minimize } & \frac{1}{2} \| \tilde{w} \|_2^2 + \frac{C}{2} \sum_{i=1}^{n} \xi_i^2 - \rho \\
\text{subject to } & y_i \tilde{w}^T (x_i \odot d) \geq \rho - \xi_i, \ i = 1, \ldots, n,
\end{align*}
\]

Note in general $B \ll m$, we obtain a new sparse model:

\[
\text{minimize } \min_{d \in \mathcal{D}} \min_{\tilde{w}, \xi, \rho} \frac{1}{2} \| \tilde{w} \|_2^2 + \frac{C}{2} \sum_{i=1}^{n} \xi_i^2 - \rho
\]

where $d \in \mathcal{D} = \{ \| d \|_0 \leq B, \ d_j \in \{0, 1\}, \ j = 1, \ldots, m \}$. 

Mingkui Tan, Li Wang, Ivor W. Tsang
ICML2010, Haifa, Israel, June, 20-24, 2010
**New Sparse $l_2$-SVM Model**

- Original SSVM model:

$$\min_{d \in D} \min_{\tilde{w}, \xi, \rho} \frac{1}{2} \|\tilde{w}\|_2^2 + \frac{C}{2} \sum_{i=1}^{n} \xi_i^2 - \rho$$  \hspace{1cm} (3.2)

s.t. \hspace{1cm} \text{y}_i \tilde{w}'(x_i \odot d) \geq \rho - \xi_i, \hspace{1cm} i = 1, \ldots, n,

- Dual presentation:

The inner minimization problem can be solved by its dual:

$$\min_{d \in D} \max_{\alpha \in \mathcal{A}} - \frac{1}{2} \left\| \sum_{i=1}^{n} \alpha_i y_i (x_i \odot d) \right\|^2 - \frac{1}{2C} \alpha' \alpha,$$  \hspace{1cm} (3.3)

where $\alpha$ is dual variable and $\mathcal{A} = \{ \alpha \mid \sum_{i=1}^{n} \alpha_i = 1, \alpha_i \geq 0 \}$.

- This is still a mixed integer programming (MIP) problem.
Inspired by [Li et al. (2009b)], this MIP problem can be relaxed as a convex QCQP problem:

\[
\max_{\alpha \in A, \theta} -\theta \quad (3.4)
\]

\[
\theta \geq -\frac{1}{2} \| \sum_{i=1}^{n} \alpha_i y_i (x_i \odot d) \|^2 - \frac{1}{2C} \alpha' \alpha, \quad \forall \: d^t \in D
\]

- The number of \( d_t \) in \( D \) is as much as \( O(\sum_{i=0}^{r} C^i_m) \)! Hence the number of constraints is exponential and nearly infinite when \( m \) becomes large!
- However, only a few constraints are active [Li et al. (2009b)]!
MKL Formulation

By applying Lagrangian theory, we arrive its dual form:

\[
\min_{\mu \in \mathcal{M}} \max_{\alpha \in \mathcal{A}} -\frac{1}{2} (\alpha \odot y)' \left( \sum_{d^t \in \mathcal{D}} \mu_t X_t X'_t + \frac{1}{C} I \right) (\alpha \odot y)
\]
\[\text{s.t. } \sum \mu_t = 1, \mu_t \geq 0\]

(3.5)

where \( X_t = [x_1 \odot d^t, \ldots, x_n \odot d^t]' \).

- This problem is a multiple kernel learning (MKL) problem and each \( X_t X'_t \) (determined by \( d^t \)) defines one base kernel.
- The base kernels in (3.5) are exponential and nearly infinite when \( m \) becomes large!
- Only a few kernels are effective!
**FGM with Cutting Plane Algorithm**

**Challenge:** How to solve such *nearly infinite kernel learning* problem?

- Inspired by infinite kernel learning [Gehler et al.(2008)], we use *cutting plane algorithm* to solve it.
- It *iteratively* finds the most violated constraint $d^t$ and adds it into the *active constraint set* $C$.
- After each $d^t$ is obtained, the best combination will be learned by MKL.

As the proposed method *iteratively “generates”* the most informative features (indexed by $d^t$), we call it as Feature Generating Machine (FGM). The whole scheme is shown in the next figure.
FGM with Cutting Plane Algorithm

\[ x \odot d = \begin{cases} 
  x & \text{if } d_i = 0 \\
  d_i & \text{if } d_i = 1 
\end{cases} \]

\[ x \odot d = \begin{array}{ccccccc}
  \text{d_0} & \text{d_{t-1}} & \text{d_t} \\
  0 & 0 & 1 & 0 & 1 & 0 & 1 \\
\end{array} \]

Relaxation

\[ x \odot d_i = \left\{ \begin{array}{c}
  d_0 \\
  \vdots \\
  d_{t-1} \\
  d_t \\
\end{array} \right\} \]

\[ \alpha_{t-1}, \mu_{t-1} \]

\[ \text{Find } d_t \]

\[ \max_d \left\| \sum_{i=1}^{*} \alpha_i y_i (x_i \odot d_i) \right\|^2 \]

Mingkui Tan, Li Wang, Ivor W. Tsang

ICML2010, Haifa, Israel, June, 20-24, 2010
Two key problems remains

- How to solve the MKL subproblem?
  We use simpleMKL algorithm [Rakotomamonjy et al.(2008)].
- How to find the most violated $d^t$?
  - Recall the QCQP problem is:
    \[
    \max_{\alpha \in A, \theta} -\theta : \theta \geq -\frac{1}{2} \| \sum_{i=1}^{n} \alpha_i y_i (x_i \odot d) \|^2 - \frac{1}{2C} \alpha' \alpha, \quad \forall d^t \in D
    \]
  - Given $\alpha$, to find the most violated constraint, we need to solve:
    \[
    \max_{d \in D} \frac{1}{2} \| \sum_{i=1}^{n} \alpha_i y_i (x_i \odot d) \|^2 = \max_{d \in D} \frac{1}{2} \sum_{j=1}^{m} c_j^2 d_j
    \]
    with $c_j = \sum_{i=1}^{n} \alpha_i y_i x_{ij}$.
    - It is a linear integer optimization problem under $\sum_{j=1}^{m} d_i \leq B$.
    - It can be easily and globally solved by first sorting $c_j^2$s and then setting the first $B$ $d_j$ to 1 and the rests to 0.
  - The index $c_j^2$ can be considered as the feature score to features.
Global Convergence

Theorem

Assume that the sub-problem of MKL and the most violated $d^t$ selection in step 3 can be exactly solved, FGM can **globally converge** after a finite number of steps.

- By proving that FGM can monotonically improve the objective values of the QCQP problem which is upper bounded, we can complete the proof [Chen & Ye(2008)].

- Why FGM sparse?
  Empirically, FGM stops no more than 10 iterations. Hence no more than 10 $d_t$ will be obtained. Then the final feature is no more than $B_1 = 10B \ll m$. **FGM is sparse!**

- By varying $B$, we can easily control the sparsity!
FGM scales linearly in computation with respect to $n$ and $m$. In other words, FGM takes $O(mn)$ time complexity.

- FGM only needs a small number ($\ll m$) of SVM trainings. Further we use Liblinear as our SVM solver which scales linearly with respect to $n$ and $m$ [Hsieh et al. (2008)]. Then FGM takes $O(mn)$.
- SVM-RFE takes $O(m^2n)$ when chunk size $= 1$.
- NMMKL takes $O(m^2n)$.
- QCQP-SSVM and SDP-SSVM are even more expensive!
In toy experiments, we use two groups of features, as shown above.

- For non-monotonic feature selection, one needs to select $f_1$ and $f_2$ as the most informative features.
- Use FGM-B to denote the results by using the best $B$ features of FGM.
- Use SVM(Ideal) to denote the ideal results by using $f_1$ and $f_2$. 
We gradually increase the number of noise features and test whether the considered methods can identify $f_1$ and $f_2$ when $B = 2$.

Only FGM-B can obtain the same results as the SVM(IDEAL) (indicates FGM also can). So, FGM is “non-monotonic”.

FGM shows competitive scalability regarding increasing features!
Large Scale Real-data Experiments: Datasets

**Table:** Large-scale datasets used in the experiments

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Features</th>
<th># Training pts.</th>
<th># Test pts.</th>
</tr>
</thead>
<tbody>
<tr>
<td>real-sim</td>
<td>20,958</td>
<td>32,309</td>
<td>40,000</td>
</tr>
<tr>
<td>rcv1.binary</td>
<td>47,236</td>
<td>20,242</td>
<td>677,399</td>
</tr>
<tr>
<td>news20.binary</td>
<td>1,355,191</td>
<td>9,996</td>
<td>10,000</td>
</tr>
<tr>
<td>URL0</td>
<td>3,231,961</td>
<td>16,000</td>
<td>20,000</td>
</tr>
<tr>
<td>URL1</td>
<td>3,231,961</td>
<td>20,000</td>
<td>20,000</td>
</tr>
</tbody>
</table>

- These datasets are with huge dimensions and large number of instances!
- Only FGM and SVM-RFE can work on such large datasets!
FGM (FGM-B) get better prediction accuracy on large datasets!

We can easily control the sparsity of FGM by varying $B$!
For SVM-RFE, we use very large chunk size to speed up the process.

FGM shows better scalability than SVM-RFE on large datasets on training time even with large chunk size!
Conclusions

- A feature generating machine (FGM) is proposed to learn the sparsity of input features.
- FGM can globally converge within a small number of iterations and therefore preserves the “non-monotonic” property.
- FGM scales linearly both on dimensions and instances.
- Empirically, FGM shows great scalability on large-scale and very high dimensional problems.
Thank you!
Blum, A. L. and Langley, P.
Selection of relevant features and examples in machine learning.

Bradley, P. S. and Mangasarian, O. L.
Feature selection via concave minimization and support vector machines.

Chan, A.B., Vasconcelos, N., and Lanckriet, G.R.G.
Direct convex relaxations of sparse SVM.

Chen, J. and Ye, J.
Training SVM with indefinite kernels.

Fung, G.M. and Mangasarian, O.L.
A feature selection newton method for support vector machine classification.


Guyon, I. and Elisseeff, A.

An introduction to variable and feature selection.


Guyon, I., Weston, J., Barnhill, S., and Vapnik, V.

Gene selection for cancer classification using support vector machines.


Hsieh, C.-J., Chang, K.-W., Lin, C.-J., Keerthi, S. S., and Sundararajan, S.

A dual coordinate descent method for large-scale linear SVM.


Joachims, T.
Training linear SVMs in linear time.
In *ACM KDD*, 2006.

Kelley, J. E.
The cutting plane method for solving convex programs.

Kim, S.-J. and Boyd, S.
A minimax theorem with applications to machine learning, signal processing, and finance.

Li, Y.F., Kwok, J.T., Tsang, I.W., and Zhou, Z.H.
A convex method for locating regions of interest with multi-instance learning.
In *ECML*, 2009a.

Li, Y.F., Tsang, I.W., Kwok, J.T., and Zhou, Z.H.
Tighter and convex maximum margin clustering.
In *AISTATS*, 2009b.

Ma, J., Saul, L. K., Savage, S., and Voelker, G. M.  
Identifying suspicious URLs: An application of large-scale online learning.  
In *ICML*, 2009.

Rakotomamonjy, A., F., Bach, Y., Grandvalet, and S., Canu.  
SimpleMKL.  

Weston, J., Elisseeff, A., and Scholkopf, B.  
Use of the zero-norm with linear models and kernel methods.  

Non-monotonic feature selection.  
In *ICML*, 2009.

Zhu, J., Rossett, S., Hastie, T., and Tibshirani, R.  
1-norm support vector machines.

P. Gehler and S. Nowozin.
Infinite kernel learning.

A comparison of optimization methods and software for large-scale $l_1$-regularized linear classification.
Technical report, Department of Computer Science and and Information Engineering, National Taiwan University, 2009.