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

Outline

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.

l_p Regularization

Given $\{x_i, y_i\}_{i=1}^n, x_i \in R^m$, l_p regularization minimizes the structural risk functional:

$$\min_w \Omega(\|w\|_p) + C \sum_{i=1}^{n} \ell(-y_i w' x_i), \tag{1.1}$$

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:
  - AROM[Weston et al.(2003)]; QCQP-SSVM & SDP-SSVM [Chan et al.(2007)], ...
  - Too relax for l_0-norm.
  - Computationally expensive.

l_p Regularization

When $p = 1$, it solves a linear programming problem (LPSVM) [Bradley & Mangasarian(1998), Fung & Mangasarian(2004)].

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

- 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 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.

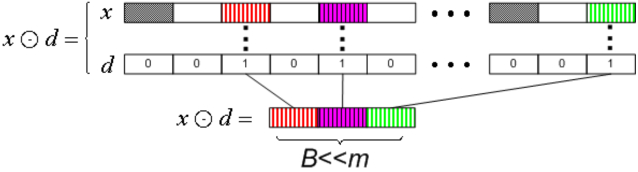
SVM-RFE is used as the baseline method in our experiments.

Contributions

- 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:



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

$$\min_{d \in \mathcal{D}} \min_{\tilde{w}, \xi, \rho} \quad \frac{1}{2}\|\tilde{w}\|_2^2 + \frac{C}{2}\sum_{i=1}^{n} \xi_i^2 - \rho \tag{3.1}$$

$$s.t. \quad y_i \tilde{w}'(x_i \odot d) \geq \rho - \xi_i, \ \ i = 1, \cdots, n,$$

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

New Sparse l_2-SVM Model

- Original SSVM model:

$$\begin{aligned} \min_{d \in \mathcal{D}} \min_{\tilde{w}, \xi, \rho} \quad & \frac{1}{2}\|\tilde{w}\|_2^2 + \frac{C}{2}\sum_{i=1}^{n} \xi_i^2 - \rho \\ s.t. \quad & y_i \tilde{w}'(x_i \odot d) \geq \rho - \xi_i, \;\; i = 1, \cdots, n, \end{aligned} \tag{3.2}$$

- Dual presentation:
  The inner minimization problem can be solved by its dual:

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

  where α is dual variable and $\mathcal{A} = \left\{ \alpha \middle| \sum_{i=1}^{n} \alpha_i = 1, \alpha_i \geq 0 \right\}$.
- This is still a mixed integer programming (MIP) problem.

Mild Convex Relaxation

Inspired by [Li et al.(2009b)], this MIP problem can be relaxed as a convex QCQP problem:

$$\max_{\alpha \in \mathcal{A}, \theta} -\theta \tag{3.4}$$

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

- The number of d_t in $\mathcal{D}$ is as much as $O(\sum_{i=0}^{r} C_m^i)$! 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:

$$\min_{\mu \in \mathcal{M}} \max_{\alpha \in \mathcal{A}} -\frac{1}{2}(\alpha \odot y)'\Big(\sum_{d^t \in \mathcal{D}} \mu_t X_t X_t' + \frac{1}{C} I\Big)(\alpha \odot y) \\ s.t. \sum \mu_t = 1, \mu_t \geq 0 \tag{3.5}$$

where $X_t = [x_1 \odot d^t, \cdots, x_n \odot d^t]'$.

- 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!

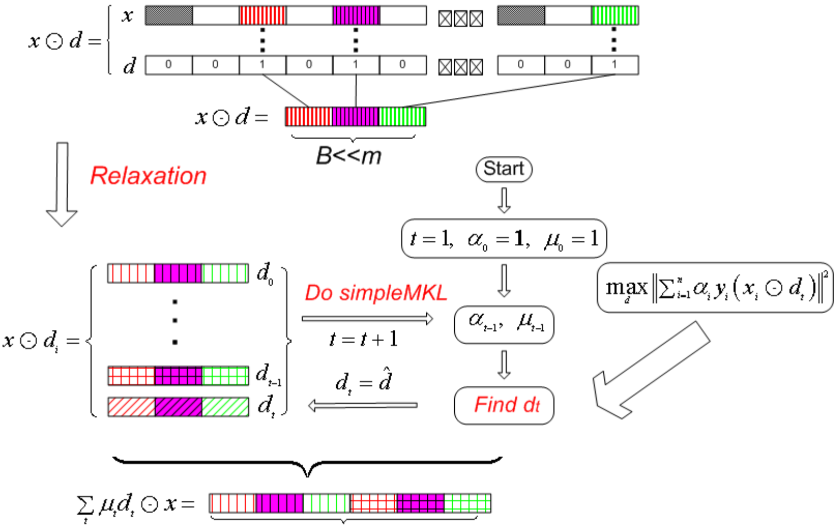
FGM with Cutting Plane Algorithm

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 $\mathcal{C}$.
- 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



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:

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

  - Given α, to find the most violated constraint, we need to solve:

$$\max_{d\in\mathcal{D}} \ \frac{1}{2}\left\|\sum_{i=1}^{n}\alpha_i y_i(x_i \odot d)\right\|^2 = \max_{d\in\mathcal{D}} \frac{1}{2}\sum_{j=1}^{m} c_j^2 d_j$$

  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^2s 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!

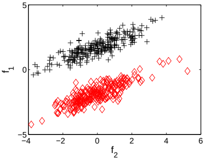
Complexity Analysis

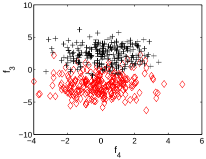
Theorem

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!

Toy Experiments

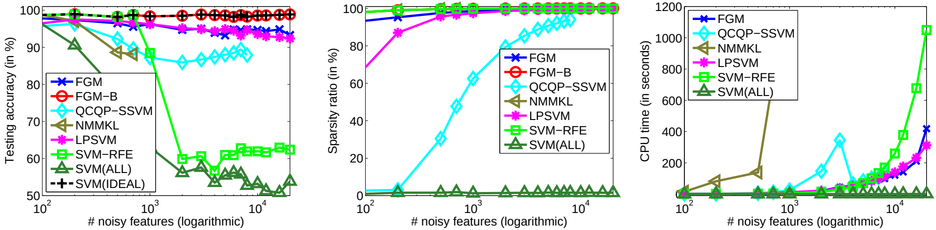




In toy experiments, we use two groups of features, as shown as above.

- 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 the best B features of FGM.
- Use SVM(IDEAL) to denote the ideal results by using f_1 and f_2.

Toy Experiments: Varying Noise Features



- 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!

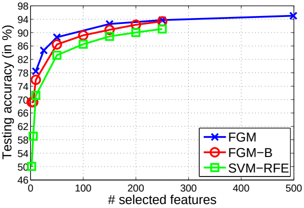
Large Scale Real-data Experiments: Datasets

Table: Large-scale datasets used in the experiments

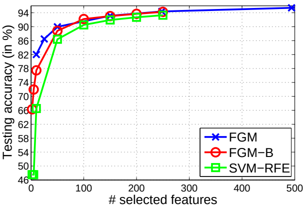
Dataset	# Features	# Training pts.	# Test pts.
real-sim	20,958	32,309	40,000
rcv1.binary	47,236	20,242	677,399
Arxiv astro-ph	99,757	62,369	32,487
news20.binary	1,355,191	9,996	10,000
URL0	3,231,961	16,000	20,000
URL1	3,231,961	20,000	20,000

- These datasets are with huge dimensions and large number of instances!
- Only FGM and SVM-RFE can work on such large datasets!

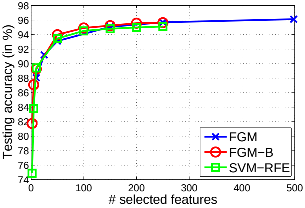
Large Scale Experiments: Accuracy



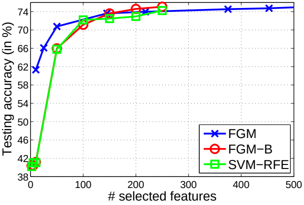
real-sim



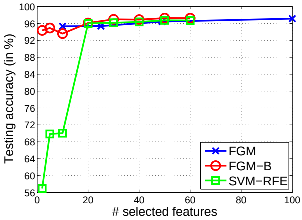
rcv1



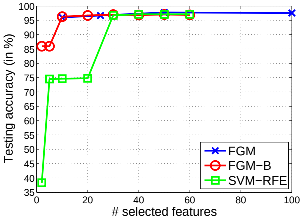
physic



news20



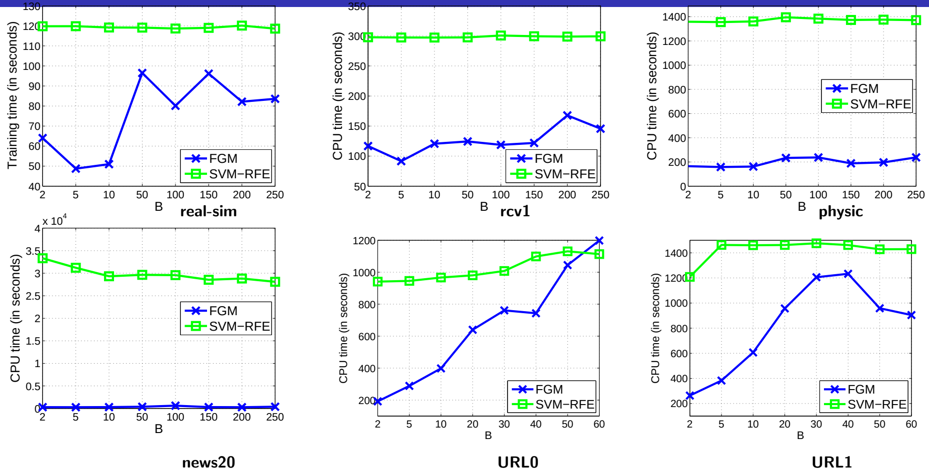
URL0



URL1

- FGM(FGM-B) get better prediction accuracy on large datasets!
- We can easily control the sparsity of FGM by varying B!

Large Scale Experiments: Training time



- 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.
Artificial Intelligence, 97:245–271, 1997.

Bradley, P. S. and Mangasarian, O. L.
Feature selection via concave minimization and support vector machines.
In *ICML*, 1998.

Chan, A.B., Vasconcelos, N., and Lanckriet, G.R.G.
Direct convex relaxations of sparse SVM.
In *ICML*, 2007.

Chen, J. and Ye, J.
Training SVM with indefinite kernels.
In *ICML*, 2008.

Fung, G.M. and Mangasarian, O.L.

A feature selection newton method for support vector machine classification.
Computational Optimization and Applications, 28:185–202, 2004.

Guyon, I. and Elisseeff, A.
An introduction to variable and feature selection.
J. Mach. Learn. Res., 3:1157–1182, 2003.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V.
Gene selection for cancer classification using support vector machines.
Machine Learning, 46:389–422, 2002.

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.
In *ICML*, 2008.

Joachims, T.

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

Kelley, J. E.
The cutting plane method for solving convex programs.
Journal of the Society for Industrial and Applied Mathematics, 8(4):703–712, 1960.

Kim, S.-J. and Boyd, S.
A minimax theorem with applications to machine learning, signal processing, and finance.
SIAM Journal on Optimization, 2008.

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.
J. Mach. Learn. Res., 9:2491–2521, 2008.

Weston, J., Elisseeff, A., and Scholkopf, B.
Use of the zero-norm with linear models and kernel methods.
J. Mach. Learn. Res., 3:1439–1461, 2003.

Xu, Z., Jin, R., J., Ye, Lyu, Michael R., and I, King.
Non-monotonic feature selection.
In *ICML*, 2009.

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

In *NIPS*, 2003.

P. Gehler and S. Nowozin.
Infinite kernel learning.
Technical Report TR-178, Max Planck Institute for Biological Cybernetics, 2008.

G.-X. Yuan, K.-W. Chang, C.-J. Hsieh, and C.-J. Lin.
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.