Learning SVM Classifiers with Indefinite Kernels

Suicheng Gu and **Yuhong Guo**

Dept. of Computer and Information Sciences

Temple University

Support Vector Machines (SVMs)

- (Kernel) SVMs are widely used in various learning scenarios, due to
  - Nice theoretical properties.
  - Good generalization performance!

Support Vector Machines (SVMs)

- Dual formulation of standard SVMs:

$$\max_{\alpha} \quad \alpha^\top e - \frac{1}{2}\alpha^\top Y K_0 Y \alpha$$
$$\text{s.t.} \quad \alpha^\top \text{diag}(Y) = 0, \ \ 0 \leq \alpha \leq C$$

  - A natural form to address nonlinear classification using kernels: $K_0 = \Phi^\top \Phi$
  - Efficient global training using convex quadratic solvers.
  - Natural on data expressed using pairwise similarities! (on conditions)

Problem

- Standard SVMs require positive semi-definite property of the kernel matrix K_0

Problem

- In many applications, pairwise similarity is a natural, convenient or suitable way of data expression.
- But the underlying similarity functions produce indefinite kernel matrices
  - E.g., similarity matrix produced by protein sequence similarity measures; by KL-divergence between probability distributions
  - syntactic kernels are shown to be useful for automatic relational learning from pairs of natural language sentences [A. Moschitti, F. Zanzotto, ICML07]

Question

- Can we still apply SVMs with indefinite kernels?
  - Not directly, but yes …

Methods

Given indefinite kernel matrix $K_0 = U \Lambda U^\top$

$$\Lambda = diag(\lambda_1, \ldots, \lambda_N)$$

- Simple spectrum modification methods:
  - **Clip:** drop all negative eigenvalues

$$K_{clip} = U\mathrm{diag}(\max(\lambda_1, 0), \cdots, \max(\lambda_N, 0))U^\top.$$

  - **Flip:** flip the sign of negative eigenvalues

$$K_{flip} = U\mathrm{diag}(|\lambda_1|, \cdots, |\lambda_N|)U^\top$$

  - **Shift:** shift the whole spectrum to remove negative eigenvalues

$$K_{shift} = U\mathrm{diag}(\lambda_1 + \eta, \ldots, \lambda_N + \eta)U^\top$$

Methods

Given indefinite kernel matrix $K_0 = U \Lambda U^\top$
$\Lambda = diag(\lambda_1, \ldots, \lambda_N)$

- Simple spectrum modification methods:
  - **Clip:**
  - **Flip:**
  - **Shift:**

Straightfoward and simple to use

Changed data independent of the classification.

Information valuable for classification might be lost

Methods

- Learn approximated p.s.d. kernel matrix and simultaneously train the classification model

  (Chen and Ye 2008; Chen, Gupta, and Recht 2009; Luss and d'Aspremont 2007)

  - E.g.,

$$\begin{aligned} \max_{\alpha} \min_{K} \quad & \alpha^\top e - \frac{1}{2}\alpha^\top Y K Y \alpha + \rho \|K - K_0\|_F^2 \\ \text{s.t.} \quad & \alpha^\top \text{diag}(Y) = 0; \quad 0 \le \alpha \le C; \quad K \succeq 0 \end{aligned}$$

- How about the testing procedure?
  - Use the original similarities? **Inconsistent treatment of training and test samples.**
  - Solving extra large positive semi-definite programming ? **Provide some consistency, but with computational cost, not a principled solution**

Proposed Approach

- A novel joint optimization model over SVMs and kernel principal component analysis (KPCA) for learning with indefinite kernels
  - reformulate the KPCA into a general kernel transformation framework
  - Incorporate the framework into SVM classifications to formulate a joint convex optimization problem
  - Principled and consistent transformations over training and test samples

KPCA Framework

- Given high-dimensional feature map ϕ of data X, $K_0 = \Phi^\top \Phi$, KPCA minimizes the reconstruction loss, transform the data to low dimension $Z = W^\top \Phi$

$$\min_W \left\|\Phi - WW^\top \Phi\right\|_F^2, \quad \text{s.t. } W^\top W = I_d$$

Drop Φ

$$\max_V \; tr(V^\top K_0 K_0 V), \quad \text{s.t. } V^\top K_0 V = I_d.$$

$$K_v = Z^\top Z = \Phi^\top W W^\top \Phi = K_0 V V^T K_0.$$

Kernel transformation

KPCA Framework

- Generalization of the kernel transformation:

$$\max_{V} tr(V^\top K_0 K_0 V), \quad \text{s.t. } V^\top K_0 V = I_d.$$

works for indefinite kernel matrix as well, as long as a feasible *d* value is given and V has real values

- Principled and consistent transformations:
  - on training samples: $K_v = K_0 V V^T K_0$
  - for a similarity vector between training samples and a new test sample x:

$$k_v = K_0 V V^T k_0$$

Original similarities

Connections with Spectrum Modifications

- Using different **V** matrix, the transformation framework $K_v = K_0 V V^T K_0$ can recover the spectrum modification methods:
  - **Clip:**

$$V_{clip} = U|\Lambda|^{-\frac{1}{2}} \mathrm{diag}\left(I_{\{\lambda_1>0\}}, \ldots, I_{\{\lambda_N>0\}}\right)$$

  - **Flip:**

$$V_{flip} = U|\Lambda|^{-\frac{1}{2}}$$

  - **Shift:**

$$V_{shift} = U|\Lambda|^{-1}(\Lambda + \eta I)^{\frac{1}{2}}$$

Training SVM with Indefinite Kernels

- A joint optimization over SVM and KPCA

$$\min_{\mathrm{w},b,\xi,V} \quad \frac{1}{2}\mathrm{w}^\top \mathrm{w} + C\sum_i \xi_i - \rho\, tr(V^\top K_0 K_0 V)$$

$$\text{s.t.} \quad y_i(\mathrm{w}^\top V^\top K_0(:,i) + b) \geq 1 - \xi_i, \;\; \xi_i \geq 0, \; \forall i;$$

$$V^\top K_0 V = I_d; \;\; K_0 V V^\top K_0 \succeq 0.$$

$\left\|\Phi - WW^\top\Phi\right\|_F^2$ Distance regularization in the feature space

- Alternatively, consider Dual SVM:

$$\max_{\alpha} \min_{V} \quad \alpha^\top e - \frac{1}{2}\alpha^\top Y K_0 V V^T K_0 Y \alpha$$

$$-\rho\, tr(V^\top K_0 K_0 V)$$

$$\text{s.t.} \quad \alpha^\top \mathrm{diag}(Y) = 0; \;\; 0 \leq \alpha \leq C;$$

$$V^\top K_0 V = I_d;$$

Convex!

For 2-class SVMs

Multiclass SVMs

- 1-vs-1 strategy
  - Training multiple binary SVMs independently?

    Problem: different transformation matrix **V** can be learned for each pair of classes, which leads to inconsistent transformation of the training samples.
  - **A joint training framework:**

    Idea: maintain one overall kernel transformation

$$K_v = K_0 V V^T K_0$$

the kernel matrix involved in a pair of classes a,b is a sub-matrix of K_v by selecting related entries

$$K_{ab} = D_{ab}^{\top} K_v D_{ab}.$$

Multiclass SVMs

- 1-vs-1 strategy: **A joint training framework:**

$$\max_{\alpha} \min_{V} \sum_{1\le a<b\le k} \left(\alpha_{ab}^{\top} e - \frac{1}{2} \alpha_{ab}^{\top} Y_{ab} D_{ab}^{T} K_0 V V^{T} K_0 D_{ab} Y_{ab} \alpha_{ab} \right) - \rho \, tr(V^{\top} K_0 K_0 V)$$

$$\text{s.t.} \quad \alpha_{ab}^{\top} \text{diag}(Y_{ab}) = 0, \;\; \forall 1 \le a < b \le k;$$
$$0 \le \alpha_{ab} \le C, \;\; \forall 1 \le a < b \le k;$$
$$V^{\top} K_0 V = I_d$$

A convex optimization problem.

- Optimization: alternative optimization procedure

Experiments

- Synthetic Experiments
  - Constructed four 3-class data sets
  - Each data set is generated using three Gaussian distributions with covariance matrix $\Lambda = \text{diag}(\sigma^2, \sigma^2)$ and mean vectors $\mu_1 = (-3, 3)$, $\mu_2 = (3, -3)$ and $(3\sqrt{3}, 3\sqrt{3})$
  - Add Gaussian noise to the linear kernel matrix to produce **indefinite** kernel matrix

$$K_0(i, j) = \mathrm{x}_i^T \mathrm{x}_j + z_{ij}, \text{ where } z_{ij} \sim N(0, \eta).$$

- Eight real-world data sets with indefinite kernels produced by different similarity measures

Synthetic Experiments

Characteristics of the four synthetic data sets

Data	σ^2	η	λ_{min}	$\left\|\frac{\lambda_{min}}{\lambda_{max}}\right\|$	$\left\|\frac{\sum \lambda_i^-}{\sum \lambda_j^+}\right\|$
Synth 1	2	20	-143	.02	.47
Synth 2	2	100	-693	.11	.82
Synth 3	4	20	-140	.02	.44
Synth 4	4	100	-702	.11	.80

classification errors (%)

Data	Clip	Flip	Shift	Robust SVM	IKFD	SVM-CA
Synth 1	1.50	2.00	15.83	1.53	1.20	0.72
Synth 2	9.67	11.00	22.33	9.05	2.43	1.83
Synth 3	4.00	4.83	21.50	4.11	1.69	1.17
Synth 4	16.17	16.67	38.17	15.24	4.70	3.50

Synthetic Experiments

Characteristics of the four synthetic data sets

Data	σ^2	η	λ_{min}	$\left\lvert\frac{\lambda_{min}}{\lambda_{max}}\right\rvert$	$\left\lvert\frac{\sum \lambda_i^-}{\sum \lambda_j^+}\right\rvert$
Synth 1	2	20	-143	.02	.47
Synth 2	2	100	-693	.11	.82
Synth 3	4	20	-140	.02	.44
Synth 4	4	100	-702	.11	.80

classification errors (%)

Data	Clip	Flip	Shift	Robust SVM	IKFD	SVM-CA
Synth 1	1.50	2.00	15.83	1.53	1.20	0.72
Synth 2	9.67	11.00	22.33	9.05	2.43	1.83
Synth 3	4.00	4.83	21.50	4.11	1.69	1.17
Synth 4	16.17	16.67	38.17	15.24	4.70	3.50

Real World Data Sets

classification error rates (%) on binary classification data sets. '

Dataset	Yeast5v7	Yeast5v12	Yeast7v12	Amazon	Aural Sonar	Voting
Clip+SVM	40.0±1.1	20.0±1.3	25.5±1.2	10.3±0.9	11.2±0.8	3.0±0.3
Flip+SVM	46.0±0.6	17.8±1.2	22.0±1.0	11.0±0.9	16.8±0.9	3.2±0.3
Shift+SVM	35.0±0.5	42.8±1.5	46.7±1.9	16.0±0.8	17.3±0.9	5.8±0.5
IKFD	34.2±1.0	17.5±1.0	14.0±1.0	15.6±0.9	**8.4±0.6**	5.7±0.3
Robust SVM	29.0±1.0	18.0±1.0	15.0±0.9	**8.8±0.8**	11.0±0.9	3.3±0.3
SVM-CA	**25.0±0.9**	**10.7±0.8**	**10.5±0.8**	9.5±0.9	8.6±0.6	**2.7±0.3**

classification error rates (%) on multi-class classification data sets.

Dataset	Protein	Glass	Patrol	Catcortex
Clip+SVM	6.3±0.7	41.1±1.2	48.6±1.5	10.5±2.0
Flip+SVM	4.0±0.7	39.4±1.1	44.8±1.4	13.5±2.3
Shift+SVM	5.5±0.7	38.3±0.9	51.4±1.5	49.0±4.0
IKFD	8.2±0.9	43.3±1.1	25.7±1.8	12.5±1.9
Robust SVM	16.4±1.1	39.1±1.0	31.3±1.4	9.4±1.7
SVM-CA	**2.5±0.5**	**37.3±0.8**	**12.4±0.8**	**4.5±1.4**

Thanks!