Feature Selection with the Potential Support Vector Machine

Sepp Hochreiter

Technische Universität Berlin
Fakultät für Elektrotechnik und Informatik

NIPS Workshop Feature Extraction/Selection 12.12.2003
Non-linear Support Vector Machine

- class -1
- class +1
- support vectors

NIPS Workshop Feature Extraction/Selection 12.12.2003
Non-linear Support Vector Machine

- class −1
- class +1
- support vectors
Non-linear Support Vector Machine

Kernel Trick

„Kernel Trick“ replaces the dot product with kernel $k$:

$k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ (\(\phi\) may be unknown)

Kernel matrix $K$, $K_{i,j} = k(x_i, x_j)$, is sufficient for model selection.
Basic Idea for Feature Selection

- class -1
- class +1
- feature objects
- support vectors

NIPS Workshop Feature Extraction/Selection 12.12.2003
Basic Idea for Feature Selection

Consequences

- Two sets of objects: objects to classify and complex feature objects
- Both object sets are mapped into the same space (2 mappings)
- Expansion of the normal vector with respect to the feature objects
  - feature weighting (SVM weights the objects to classify)
  - feature extraction

Problem

- Features are associated with vectors: which vectors?

Solution

- Data matrix (objects x features) is dot product matrix between complex feature vectors and object vectors
Basic Idea for Feature Selection

Problems with data matrix as dot product matrix

- Kernel matrix $K$ may be not positive definite and not squared (SVM optimization is not possible)
- Dot products between objects to classify are unknown
- SVM technique cannot be applied because $\|w\|$ cannot be computed

New objective necessary
New constraints necessary
Scale Invariant Objective

SVM solution and error bounds depend on scaling
Scale Invariant Objective

\[ \begin{align*} d_x & \quad d_x \\ d_y & \quad d_y \\ R & \quad \tilde{R} \end{align*} \]

Scale invariant objective derived from covering number error bounds:

New: \[ \left\| X^T w \right\|_2^2 \]

SVM: \[ \left\| w \right\|_2^2 \]

\[ \left\| X^T w \right\|_2^2 = w^T X X^T w \] (\(X\) is matrix of vectors \(x^i\))

NIPS Workshop Feature Extraction/Selection 12.12.2003
New Constraints

Residual error w.r.t. classification function:

\[ r_\alpha = (w^T \cdot x^\alpha) + b - y^\alpha \]

Minimize the quadratic loss function:

\[ R_{emp} = \frac{1}{2p} \sum_{\alpha=1}^{p} r_\alpha^2 \overset{!}{=} \min \]

\[ p \nabla_w R_{emp} = X(X^T w + b1 - y) = 0 \]

for a linear classifier
New Constraints

Derivative of $R_{\text{emp}}$ with respect to $w$ along direction $z_j$ should be zero:

$$\frac{dR_{\text{emp}}(w + tz_j)}{dt} = z_j^T \nabla_w R_{\text{emp}} = z_j^T \sum_{\alpha=1}^{p} r_{\alpha} x_{\alpha} = 0$$

- $w$ takes on $R_{\text{emp}}$'s minimum along $z_j$
- $z_j$ are the complex feature vectors

All directional constraints in matrix form

$$K^T(X^T w + b1 - y) = 0$$

where

$$K = X^T Z$$

Number of constraints is now the number of complex features
New Constraints

Measurement noise: constraints may lead to overfitting

\[ K^T (X^T w + b1 - y) - \epsilon \leq 0 \]
\[ K^T (X^T w + b1 - y) + \epsilon \geq 0 \]

Normalization of $K$ to equal feature variance is necessary to use one $\epsilon$ for all constraints, that is for all complex features.

Increase of the residual error after the elimination of the $j$-th feature is bounded by:

\[ 2 \epsilon |w_j| + p w_j^2 \]
Potential Support Vector Machine

Algorithm

Primal

\[
\min_{w,b} \frac{1}{2} \|X^T w\|_2^2
\]

s.t.

\[
K^T \left(X^T w + b \ 1 - y\right) + \varepsilon 1 \geq 0
\]

\[
K^T \left(X^T w + b \ 1 - y\right) - \varepsilon 1 \leq 0
\]

\(y\) is vector of labels \(y_i\)

\(Z\) is matrix of feature objects \(z_j\)

Lagrangian

\(XX^T w = XX^T Z \alpha\) is assured by \(w = Z \alpha\)

\(w\) expanded with respect to features
Potential Support Vector Machine

Algorithm

Dual

\[
\min_{\alpha^+, \alpha^-} \frac{1}{2} (\alpha^+ - \alpha^-)^T K^T K (\alpha^+ - \alpha^-) - y^T K (\alpha^+ - \alpha^-) + \varepsilon 1^T (\alpha^+ + \alpha^-)
\]

s.t. \quad 1^T K (\alpha^+ - \alpha^-) = 0, \quad C 1 \geq \alpha^+, \quad \alpha^- \geq 0

\[w = Z \alpha\], where \(\alpha = \alpha^+ - \alpha^-\).

\(K^T K\) is (features x features) and optimization would be computational expensive: Sequential Minimal Optimization (SMO)
Potential Support Vector Machine

**Characteristic**

- Works with data matrix
- Feature selection: Identification of relevant features

**Applications:** Prediction of a treatment outcome based on the gene expression profile obtained from the micro array technique

- Brain tumor
- Breast cancer
Brain Tumor

Task

Brain tumor (medulloblastoma) patients respond differently to the chemotherapy and radiation

- Negative prognoses: alternative therapy or more intensive control
- Positive prognoses: toxicity of the therapy can be reduced

60 patients and 7129 genes

Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stalvitzky, D. N. Louis, J. P. Mesirov, E. S.
Lander, T. R. Golub
Prediction of central nervous system embryonal tumour outcome based on gene expression
Breast Cancer

Task

Breast cancer: the treatment is for 70-80 % of the patients not necessary to avoid metastasis

- Prediction of metastasis leads to choice of patients for therapy
- Alternative treatment and toxicity reduction

78 patients und 25000 genes


Gene expression profiling predicts clinical outcome of breast cancer

Nature 415: 530-536, 2002
Breast Cancer

Classification results

<table>
<thead>
<tr>
<th>Standard Feature Selection</th>
<th>New Method (P-SVM)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Method</strong></td>
<td><strong>F</strong></td>
</tr>
<tr>
<td>weighted voting</td>
<td>70</td>
</tr>
</tbody>
</table>

Standard feature selection with „signal-to-noise“-statistic
Feature selection history