Efficient Maximum Margin Clustering

Changshui Zhang, Bin Zhao

Dept. Automation, Tsinghua Univ.

MLA, Nov. 8, 2008
Nanjing, China
Support Vector Machine

Given $\mathcal{X} = \{x_1, \cdots, x_n\}$, $y = (y_1, \ldots, y_n) \in \{-1, +1\}^n$, SVM finds a hyperplane $f(x) = w^T \phi(x) + b$ by solving

$$\begin{align*}
    \min_{w, b, \xi_i} & \quad \frac{1}{2} w^T w + \frac{C}{n} \sum_{i=1}^{n} \xi_i \\
    \text{s.t.} & \quad y_i (w^T \phi(x_i) + b) \geq 1 - \xi_i \\
                     & \quad \xi_i \geq 0 \quad i = 1, \ldots, n
\end{align*}$$ (1)
**Maximum Margin Clustering** [Xu et. al. 2004]

**MMC** targets to find not only the optimal hyperplane \((\mathbf{w}^*, b^*)\), but also the optimal labeling vector \(\mathbf{y}^*\)

\[
\begin{align*}
\min_{\mathbf{y} \in \{-1,+1\}^n} \quad & \min_{\mathbf{w}, b, \xi_i} \quad \frac{1}{2} \mathbf{w}^T \mathbf{w} + \frac{C}{n} \sum_{i=1}^{n} \xi_i \\
\text{s.t.} \quad & y_i(\mathbf{w}^T \phi(\mathbf{x}_i) + b) \geq 1 - \xi_i \\
\quad & \xi_i \geq 0 \quad i = 1, \ldots, n
\end{align*}
\]
Representative Works

Semi-definite programming [Xu et. al. (NIPS 2004)]
- Several relaxations made
- \( n^2 \) variables in SDP
- Time complexity \( O(n^7) \)
Representative Works

Semi-definite programming [Valizadehgan and Jin (NIPS 2006)]

- Reduce number of variables from $n^2$ to $n$
- Time complexity $O(n^4)$
- Only 2-class scenario
Representative Works

Alternating optimization [Zhang et. al. (ICML 2007)]
- Involve a sequence of QPs
- Number of iterations not guaranteed theoretically
- Only 2-class scenario
Outline

1. Motivation
2. Two-Class Maximum Margin Clustering
3. Multi-Class Maximum Margin Clustering
4. Related Works
5. Conclusions
Theorem

Maximum margin clustering is equivalent to

$$\min_{w, b, \xi_i} \quad \frac{1}{2} w^T w + \frac{C}{n} \sum_{i=1}^{n} \xi_i$$

s.t. \quad |w^T \phi(x_i) + b| \geq 1 - \xi_i

$$\xi_i \geq 0 \quad i = 1, \ldots, n$$

where the labeling vector $y_i = \text{sign}(w^T \phi(x_i) + b)$. 
Problem Reformulation

Theorem

Any solution \((\mathbf{w}^*, b^*)\) to problem (4) is also a solution to problem (3) (and vice versa), with \(\xi^* = \frac{1}{n} \sum_{i=1}^{n} \xi_i^*\).

\[
\begin{align*}
\min_{\mathbf{w}, b, \xi \geq 0} & \quad \frac{1}{2} \mathbf{w}^T \mathbf{w} + C\xi \\
\text{s.t.} & \quad \forall \mathbf{c} \in \{0, 1\}^n : \\
& \quad \frac{1}{n} \sum_{i=1}^{n} c_i |\mathbf{w}^T \phi(x_i) + b| \geq \frac{1}{n} \sum_{i=1}^{n} c_i - \xi
\end{align*}
\]

(4)
Problem Reformulation

- Number of variables reduced by $2n - 1$
- Number of constraints increased from $n$ to $2^n$
- We can always find a polynomially sized subset of constraints, with which the solution of the relaxed problem fulfills all constraints from problem (4) up to a precision of $\epsilon$. 
Cutting Plane Algorithm [J. E. Kelley 1960]

- Starts with an empty constraint subset $\Omega$
- Computes the optimal solution to problem (4) subject to the constraints in $\Omega$
- Finds the most violated constraint in problem (4) and adds it into the subset $\Omega$
- Stops when no constraint in (4) is violated by more than $\epsilon$

$$\frac{1}{n} \sum_{i=1}^{n} c_i |w^T \phi(x_i) + b| \geq \frac{1}{n} \sum_{i=1}^{n} c_i - (\xi + \epsilon)$$ (5)

Changshui Zhang, Bin Zhao  
Efficient Maximum Margin Clustering
The Most Violated Constraint

Theorem

The most violated constraint could be computed as follows

\[ c_i = \begin{cases} 
1 & \text{if } |\mathbf{w}^T \phi(x_i) + b| < 1 \\
0 & \text{otherwise}
\end{cases} \] (6)

The feasibility of a constraint is measured by the corresponding value of \( \xi \)

\[
\frac{1}{n} \sum_{i=1}^{n} c_i |\mathbf{w}^T \phi(x_i) + b| \geq \frac{1}{n} \sum_{i=1}^{n} c_i - \xi
\] (7)
Enforcing the Class Balance Constraint

Enforce class balance constraint to avoid trivially “optimal” solutions

\[
\min_{w,b,\xi \geq 0} \frac{1}{2} w^T w + C\xi
\]

\[
\text{s.t. } \forall c \in \Omega: \frac{1}{n} \sum_{i=1}^{n} c_i \|w^T \phi(x_i) + b\| \geq \frac{1}{n} \sum_{i=1}^{n} c_i - \xi
\]

\[-l \leq \sum_{i=1}^{n} \left( w^T \phi(x_i) + b \right) \leq l\]

Solve non-convex optimization problem whose objective function could be expressed as a difference of convex functions

\[
\min_z f_0(z) - g_0(z) \quad (9)
\]

s.t. \( f_i(z) - g_i(z) \leq c_i \quad i = 1, \ldots, n \)

where \( f_i \) and \( g_i \) are real-valued convex functions on a vector space \( Z \) and \( c_i \in \mathcal{R} \) for all \( i = 1, \ldots, n \).
The Constrained Concave-Convex Procedure

Given an initial point $z_0$, the CCCP computes $z_{t+1}$ from $z_t$ by replacing $g_i(z)$ with its first-order Taylor expansion at $z_t$

$$
\min_z f_0(z) - T_1\{g_0, z_t\}(z)
$$

$$
s.t. f_i(z) - T_1\{g_i, z_t\}(z) \leq c_i \quad i = 1, \ldots, n
$$
Optimization via the **CCCP**

By substituting first-order Taylor expansion into problem (8), we obtain the following *quadratic programming (QP)* problem:

\[
\min_{w,b,\xi} \frac{1}{2} w^T w + C \xi \\
\text{s.t. } \xi \geq 0
\]

\[
-l \leq \sum_{i=1}^{n} (w^T \phi(x_i) + b) \leq l
\]

\[
\forall \mathbf{c} \in \Omega: \frac{1}{n} \sum_{i=1}^{n} c_i - \xi - \frac{1}{n} \sum_{i=1}^{n} c_i \text{sign}(w^T \phi(x_i) + b) \left[ w^T \phi(x_i) + b \right] \leq 0
\]
Justification of \textit{CPMMC}

\textbf{Theorem}

\textit{For any dataset } $\mathcal{X} = (x_1, \ldots, x_n)$ \textit{and any } $\epsilon > 0$, \textit{the CPMMC algorithm for maximum margin clustering returns a point } $(w, b, \xi)$ \textit{for which } $(w, b, \xi + \epsilon)$ \textit{is feasible in problem (4).}
**Theorem**

*Each iteration of CPMMC takes time $O(sn)$ for a constant working set size $|\Omega|$.*

**Theorem**

*For any $\epsilon > 0$, $C > 0$, and any dataset $X = \{x_1, \ldots, x_n\}$, the CPMMC algorithm terminates after adding at most $\frac{CR}{\epsilon^2}$ constraints, where $R$ is a constant number independent of $n$ and $s$.***
Theorem

For any dataset \( \mathcal{X} = \{x_1, \ldots, x_n\} \) with \( n \) samples and sparsity of \( s \), and any fixed value of \( C > 0 \) and \( \epsilon > 0 \), the CPMMC algorithm takes time \( O(sn) \).
For UCI digits and MNIST datasets, we give a thorough comparison by considering all 45 pairs of digits 0-9. For NC/MMC/GMMC/IterSVR, results on the digits and ionosphere data are simply copied from (Zhang et. al., 2007).
# Speed of CPMMC

<table>
<thead>
<tr>
<th>Data</th>
<th>KM</th>
<th>NC</th>
<th>GMMC</th>
<th>IterSVR</th>
<th>CPMMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digits 3-8</td>
<td>0.51</td>
<td>0.12</td>
<td>276.16</td>
<td>19.72</td>
<td>1.10</td>
</tr>
<tr>
<td>Digits 1-7</td>
<td>0.54</td>
<td>0.13</td>
<td>289.53</td>
<td>20.49</td>
<td>0.95</td>
</tr>
<tr>
<td>Digits 2-7</td>
<td>0.50</td>
<td>0.11</td>
<td>304.81</td>
<td>19.69</td>
<td>0.75</td>
</tr>
<tr>
<td>Digits 8-9</td>
<td>0.49</td>
<td>0.11</td>
<td>277.26</td>
<td>19.41</td>
<td>0.85</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>0.07</td>
<td>0.12</td>
<td>273.04</td>
<td>18.86</td>
<td>0.78</td>
</tr>
<tr>
<td>Letter</td>
<td>0.08</td>
<td>2.24</td>
<td>-</td>
<td>2133</td>
<td>0.87</td>
</tr>
<tr>
<td>Satellite</td>
<td>0.19</td>
<td>5.01</td>
<td>-</td>
<td>6490</td>
<td>4.54</td>
</tr>
<tr>
<td>Text-1</td>
<td>66.09</td>
<td>6.04</td>
<td>-</td>
<td>5844</td>
<td>19.75</td>
</tr>
<tr>
<td>Text-2</td>
<td>52.32</td>
<td>5.35</td>
<td>-</td>
<td>6099</td>
<td>16.16</td>
</tr>
</tbody>
</table>
Dataset Size $n$ vs. Speed

- **Motivation**
  - Two-Class Maximum Margin Clustering
  - Multi-Class Maximum Margin Clustering

- **Related Works**

- **Conclusions**

---

**Dataset Size $n$ vs. Speed**

![Graph showing CPU-time vs. number of samples for different datasets](image)

- **Letter**
- **Satellite**
- **Text-1**
- **Text-2**
- **MNIST-1vs2**
- **MNIST-1vs7**
- **$O(n)$**

---

**Efficient Maximum Margin Clustering**

Changshui Zhang, Bin Zhao
$\epsilon$ vs. Accuracy & Speed

(a) Clustering Accuracy

(b) CPU-Time (seconds)

Changshui Zhang, Bin Zhao
Efficient Maximum Margin Clustering
C vs. Accuracy & Speed

(a) Clustering Accuracy

(b) CPU-Time (seconds)

Changshui Zhang, Bin Zhao

Efficient Maximum Margin Clustering
Outline

1. Motivation
2. Two-Class Maximum Margin Clustering
3. Multi-Class Maximum Margin Clustering
4. Related Works
5. Conclusions
Multi-Class Support Vector Machine [Crammer & Singer 2001]

Given a point set $\mathcal{X} = \{x_1, \cdots, x_n\}$ and their labels $y = (y_1, \ldots, y_n) \in \{1, \ldots, k\}^n$, SVM defines a weight vector $w_p$ for each class $p \in \{1, \ldots, k\}$ and classifies sample $x$ by $y^* = \arg \max_{y \in \{1, \ldots, k\}} w_y^T x$ with the weight vectors obtained as

\[
\min_{w_1, \ldots, w_k, \xi} \frac{1}{2} \beta \sum_{p=1}^{k} \|w_p\|^2 + \sum_{i=1}^{n} \xi_i
\]

s.t. $\forall i = 1, \ldots, n, r = 1, \ldots, k$

$w_{y_i}^T x_i + \delta_{y_i,r} - w_r^T x_i \geq 1 - \xi_i$
Similar with the binary clustering scenario

$$\min_{y} \min_{w_1, \ldots, w_k, \xi} \left\{ \frac{1}{2}\beta \sum_{p=1}^{k} \|w_p\|^2 + \frac{1}{n} \sum_{i=1}^{n} \xi_i \right\}$$

s.t. \quad \forall i = 1, \ldots, n, \ r = 1, \ldots, k

$$w_{y_i}^T x_i + \delta_{y_i r} - w_r^T x_i \geq 1 - \xi_i$$

(13)
Theorem

\[
\begin{align*}
\min_{\mathbf{w}_1, \ldots, \mathbf{w}_k, \xi} & \quad \frac{1}{2} \beta \sum_{p=1}^{k} \| \mathbf{w}_p \|^2 + \frac{1}{n} \sum_{i=1}^{n} \xi_i \\
\text{s.t.} & \quad \forall i = 1, \ldots, n, r = 1, \ldots, k \\
& \quad \sum_{p=1}^{k} \mathbf{w}_p^T \mathbf{x}_i \prod_{q=1, q \neq p}^{k} I(\mathbf{w}_p^T \mathbf{x}_i > \mathbf{w}_q^T \mathbf{x}_i) + \prod_{q=1, q \neq r}^{k} I(\mathbf{w}_r^T \mathbf{x}_i > \mathbf{w}_q^T \mathbf{x}_i) - \mathbf{w}_r^T \mathbf{x}_i \geq 1 - \xi_i
\end{align*}
\]  

(14)

where \( I(\cdot) \) is the indicator function and the label for sample \( \mathbf{x}_i \) is determined as \( y_i = \sum_{p=1}^{k} p \prod_{q=1, q \neq p}^{k} I(\mathbf{w}_p^T \mathbf{x}_i > \mathbf{w}_q^T \mathbf{x}_i) \)
Problem (14) can be equivalently formulated as problem (15), with $\xi^* = \frac{1}{n} \sum_{i=1}^{n} \xi_i^*$. 

$$\min_{\mathbf{w}_1, \ldots, \mathbf{w}_k, \xi} \frac{1}{2} \beta \sum_{p=1}^{k} \|\mathbf{w}_p\|^2 + \xi$$

$$s.t. \forall \mathbf{c}_i \in \{\mathbf{e}_0, \mathbf{e}_1, \ldots, \mathbf{e}_k\}, \ i = 1, \ldots, n$$

$$\frac{1}{n} \sum_{i=1}^{n} \left\{ \mathbf{c}_i^T \mathbf{e} \sum_{p=1}^{k} \mathbf{w}_p^T \mathbf{x}_i z_{ip} + \sum_{p=1}^{k} \mathbf{c}_{ip} (z_{ip} - \mathbf{w}_p^T \mathbf{x}_i) \right\} \geq \frac{1}{n} \sum_{i=1}^{n} \mathbf{c}_i^T \mathbf{e} - \xi$$

where $z_{ip} = \prod_{q=1, q\neq p}^{k} l_{(\mathbf{w}_p^T \mathbf{x}_i > \mathbf{w}_q^T \mathbf{x}_i)}$ and each constraint $\mathbf{c}$ is represented as a $k \times n$ matrix $\mathbf{c} = (\mathbf{c}_1, \ldots, \mathbf{c}_n)$. 

**Theorem**

Changshui Zhang, Bin Zhao

Efficient Maximum Margin Clustering
Problem Reformulation

- Number of variables reduced by $2n - 1$
- Number of constraints increased from $nk$ to $(k + 1)^n$
- Targets to finding a small subset of constraints, with which the solution of the relaxed problem fulfills all constraints from problem (15) up to a precision of $\epsilon$. 

Changshui Zhang, Bin Zhao
Cutting Plane Algorithm [J. E. Kelley 1960, T. Joachims 2006]

- Starts with an empty constraint subset $\Omega$
- Computes the optimal solution to problem (15) subject to the constraints in $\Omega$
- Finds the most violated constraint in problem (15) and adds it into the subset $\Omega$
- Stops when no constraint in (15) is violated by more than $\epsilon$

\[
\forall c_i \in \{e_0, e_1, \ldots, e_k\}^n, \quad i = 1, \ldots, n
\]

\[
\frac{1}{n} \sum_{i=1}^{n} \left\{ c_i^T e \sum_{p=1}^{k} w_p^T x_i z_{ip} + \sum_{p=1}^{k} c_{ip} (z_{ip} - w_p^T x_i) \right\} \geq \frac{1}{n} \sum_{i=1}^{n} c_i^T e - \xi - \epsilon
\]
The Most Violated Constraint

**Theorem**

Define $p^* = \text{arg max}_p (w_p^T x_i)$ and $r^* = \text{arg max}_{r \neq p^*} (w_r^T x_i)$ for $i = 1, \ldots, n$, the most violated constraint could be calculated as follows

$$c_i = \begin{cases} 
   e_{r^*} & \text{if } (w_{p^*}^T x_i - w_{r^*}^T x_i) < 1 \\
   0 & \text{otherwise}
\end{cases}, \quad i = 1, \ldots, n \quad (17)$$
Enforcing the Class Balance Constraint

To avoid trivially “optimal” solutions

\[
\begin{align*}
\min_{\mathbf{w}_1, \ldots, \mathbf{w}_k, \xi \geq 0} & \quad \frac{1}{2} \beta \sum_{p=1}^{k} \|\mathbf{w}_p\|^2 + \xi \\
\text{s.t.} & \quad \frac{1}{n} \sum_{i=1}^{n} \left\{ \mathbf{c}_i^T \mathbf{e} \sum_{p=1}^{k} \mathbf{w}_p^T \mathbf{x}_i z_{ip} + \sum_{p=1}^{k} c_{ip} \left( z_{ip} - \mathbf{w}_p^T \mathbf{x}_i \right) \right\} \\
& \quad \geq \frac{1}{n} \sum_{i=1}^{n} \mathbf{c}_i^T \mathbf{e} - \xi, \quad \forall [\mathbf{c}_1, \ldots, \mathbf{c}_n] \in \Omega \\
& \quad -l \leq \sum_{i=1}^{n} \mathbf{w}_p^T \mathbf{x}_i - \sum_{i=1}^{n} \mathbf{w}_q^T \mathbf{x}_i \leq l, \quad \forall p, q = 1, \ldots, k
\end{align*}
\]
Optimization via the \textit{CCCP}

Calculate the subgradients

\[
\partial_{w_r} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[ \sum_{p=1}^{k} c_i^T e \sum_{p=1}^{k} w_p^T x_i z_{ip} + \sum_{p=1}^{k} c_{ip} z_{ip} \right] \right\} \bigg|_{w=w(t)}
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} c_i^T e z_{ip}^{(t)} x_i \quad \forall r = 1, \ldots, k
\]

By substituting first-order Taylor expansion into problem (18), we obtain a \textit{quadratic programming (QP)} problem.
Justification of \textit{CPM3C}

\textbf{Theorem}

\textit{For any dataset }$\mathcal{X} = (x_1, \ldots, x_n)$ \textit{and any }$\epsilon > 0$, \textit{the CPM3C algorithm returns a point }$(w_1, \ldots, w_k, \xi)$ \textit{for which }$(w_1, \ldots, w_k, \xi + \epsilon)$ \textit{is feasible.}
## Clustering Accuracy Comparison

<table>
<thead>
<tr>
<th>Data</th>
<th>KM</th>
<th>NC</th>
<th>MMC</th>
<th>CPM3C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dig 0689</td>
<td>42.23</td>
<td>93.13</td>
<td>94.83</td>
<td>96.63</td>
</tr>
<tr>
<td>Dig 1279</td>
<td>40.42</td>
<td>90.11</td>
<td>91.91</td>
<td>94.01</td>
</tr>
<tr>
<td>Cora-DS</td>
<td>28.24</td>
<td>36.88</td>
<td>-</td>
<td>43.75</td>
</tr>
<tr>
<td>Cora-HA</td>
<td>34.02</td>
<td>42.00</td>
<td>-</td>
<td>59.75</td>
</tr>
<tr>
<td>Cora-ML</td>
<td>27.08</td>
<td>31.05</td>
<td>-</td>
<td>45.58</td>
</tr>
<tr>
<td>Cora-OS</td>
<td>23.87</td>
<td>23.03</td>
<td>-</td>
<td>58.89</td>
</tr>
<tr>
<td>Cora-PL</td>
<td>33.80</td>
<td>33.97</td>
<td>-</td>
<td>46.83</td>
</tr>
<tr>
<td>WK-CL</td>
<td>55.71</td>
<td>61.43</td>
<td>-</td>
<td>71.95</td>
</tr>
<tr>
<td>WK-TX</td>
<td>45.05</td>
<td>35.38</td>
<td>-</td>
<td>69.29</td>
</tr>
<tr>
<td>WK-WT</td>
<td>53.52</td>
<td>32.85</td>
<td>-</td>
<td>77.96</td>
</tr>
<tr>
<td>WK-WC</td>
<td>49.53</td>
<td>33.31</td>
<td>-</td>
<td>73.88</td>
</tr>
<tr>
<td>20-news</td>
<td>35.27</td>
<td>41.89</td>
<td>-</td>
<td>70.63</td>
</tr>
<tr>
<td>RCVI</td>
<td>27.05</td>
<td>-</td>
<td>-</td>
<td>61.97</td>
</tr>
</tbody>
</table>
### Rand Index Comparison

<table>
<thead>
<tr>
<th>Data</th>
<th>KM</th>
<th>NC</th>
<th>MMC</th>
<th>CPM3C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dig 0689</td>
<td>0.696</td>
<td>0.939</td>
<td>0.941</td>
<td>0.968</td>
</tr>
<tr>
<td>Dig 1279</td>
<td>0.681</td>
<td>0.909</td>
<td>0.913</td>
<td>0.943</td>
</tr>
<tr>
<td>Cora-DS</td>
<td>0.589</td>
<td><strong>0.744</strong></td>
<td>-</td>
<td>0.735</td>
</tr>
<tr>
<td>Cora-HA</td>
<td>0.385</td>
<td>0.659</td>
<td>-</td>
<td><strong>0.692</strong></td>
</tr>
<tr>
<td>Cora-ML</td>
<td>0.514</td>
<td>0.720</td>
<td>-</td>
<td><strong>0.754</strong></td>
</tr>
<tr>
<td>Cora-OS</td>
<td>0.518</td>
<td>0.522</td>
<td>-</td>
<td><strong>0.721</strong></td>
</tr>
<tr>
<td>Cora-PL</td>
<td>0.643</td>
<td>0.675</td>
<td>-</td>
<td><strong>0.703</strong></td>
</tr>
<tr>
<td>WK-CL</td>
<td>0.603</td>
<td>0.602</td>
<td>-</td>
<td><strong>0.728</strong></td>
</tr>
<tr>
<td>WK-TX</td>
<td>0.604</td>
<td>0.514</td>
<td>-</td>
<td><strong>0.707</strong></td>
</tr>
<tr>
<td>WK-WT</td>
<td>0.616</td>
<td>0.581</td>
<td>-</td>
<td><strong>0.747</strong></td>
</tr>
<tr>
<td>WK-WC</td>
<td>0.581</td>
<td>0.509</td>
<td>-</td>
<td><strong>0.752</strong></td>
</tr>
<tr>
<td>20-news</td>
<td>0.581</td>
<td>0.496</td>
<td>-</td>
<td><strong>0.782</strong></td>
</tr>
<tr>
<td>RCVI</td>
<td>0.471</td>
<td>-</td>
<td>-</td>
<td><strong>0.698</strong></td>
</tr>
<tr>
<td>Data</td>
<td>KM</td>
<td>CPM3C</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>------</td>
<td>-------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dig 0689</td>
<td>34.28</td>
<td>9.66</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dig 1279</td>
<td>17.78</td>
<td>17.47</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cora-DS</td>
<td>839.67</td>
<td>35.31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cora-HA</td>
<td>204.43</td>
<td>24.35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cora-ML</td>
<td>22781</td>
<td>69.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cora-OS</td>
<td>47931</td>
<td>13.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cora-PL</td>
<td>7791.4</td>
<td>165.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WK-CL</td>
<td>672.69</td>
<td>9.534</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WK-TX</td>
<td>766.77</td>
<td>10.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WK-WT</td>
<td>4135.2</td>
<td>10.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WK-WC</td>
<td>1578.2</td>
<td>9.041</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20-news</td>
<td>2387.8</td>
<td>215.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RCVI</td>
<td>428770</td>
<td>587.9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Dataset Size $n$ vs. Speed

- For Cora & 20News:
  - Cora−DS
  - Cora−HA
  - Cora−ML
  - Cora−OS
  - Cora−PL
  - 20News

- For WebKB & RCVI:
  - WK−CL
  - WK−HA
  - WK−WT
  - WK−WC
  - RCVI

CPU−Time (seconds) vs. Number of Samples

Efficient Maximum Margin Clustering

Changshui Zhang, Bin Zhao
Motivation
Two-Class Maximum Margin Clustering
Multi-Class Maximum Margin Clustering
Related Works
Conclusions

$\epsilon$ vs. Accuracy

Changshui Zhang, Bin Zhao

Efficient Maximum Margin Clustering
Motivation
Two-Class Maximum Margin Clustering
Multi-Class Maximum Margin Clustering
Related Works
Conclusions

$\epsilon$ vs. Speed

![Graph showing CPU-time (seconds) vs. Epsilon for Cora & 20News and WebKB & RCVI datasets.](image)

- Cora & 20News:
  - WK−CL
  - WK−TX
  - WK−WT
  - WK−WC
  - RCVI
  - O($x^{-0.5}$)

- WebKB & RCVI:
  - WK−CL
  - WK−TX
  - WK−WT
  - WK−WC
  - RCVI
  - O($x^{-0.5}$)

Changshui Zhang, Bin Zhao
Efficient Maximum Margin Clustering
Semi-Supervised Support Vector Machine

Given $\mathcal{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_l, \mathbf{x}_{l+1}, \cdots, \mathbf{x}_n\}$, where the first $l$ points in $\mathcal{X}$ are labeled as $y_i \in \{-1, +1\}$ and the remaining $u = n - l$ points are unlabeled.

$$
\begin{align*}
\min_{y_{l+1}, \ldots, y_n} \min_{\mathbf{w}, b, \xi_i} & \quad \frac{1}{2} \mathbf{w}^T \mathbf{w} + \frac{C_l}{n} \sum_{i=1}^{l} \xi_i + \frac{C_u}{n} \sum_{j=l+1}^{n} \xi_j \\
\text{s.t.} & \quad y_i [\mathbf{w}^T \phi(\mathbf{x}_i) + b] \geq 1 - \xi_i, \quad \forall i = 1, \ldots, l \\
& \quad y_j [\mathbf{w}^T \phi(\mathbf{x}_j) + b] \geq 1 - \xi_j, \quad \forall j = l + 1, \ldots, n \\
& \quad \xi_i \geq 0, \quad \forall i = 1, \ldots, l \\
& \quad \xi_j \geq 0, \quad \forall j = l + 1, \ldots, n
\end{align*}
$$
Theorem

Problem (20) is equivalent to

\[
\min_{\mathbf{w}, \xi_i} \frac{1}{2} \mathbf{w}^T \mathbf{w} + \frac{C_l}{n} \sum_{i=1}^{l} \xi_i + \frac{C_u}{n} \sum_{j=l+1}^{n} \xi_j
\]

s.t. \quad \begin{align*}
    y_i[\mathbf{w}^T \phi(\mathbf{x}_i) + b] &\geq 1 - \xi_i, \quad \forall i = 1, \ldots, l \\
    |\mathbf{w}^T \phi(\mathbf{x}_j) + b| &\geq 1 - \xi_j, \quad \forall j = l + 1, \ldots, n \\
    \xi_i &\geq 0, \quad \forall i = 1, \ldots, l \\
    \xi_j &\geq 0, \quad \forall j = l + 1, \ldots, n
\end{align*}

where the labels \( y_j, j = l + 1, \ldots, n \) are calculated as \( y_j = \text{sign}(\mathbf{w}^T \phi(\mathbf{x}_j) + b) \).
Theorem

Problem (21) can be equivalently formulated as

$$\begin{align*}
\min_{\mathbf{w}, \xi \geq 0} & \quad \frac{1}{2} \mathbf{w}^T \mathbf{w} + \xi \\
\text{s.t.} & \quad \frac{1}{n} \left\{ C_l \sum_{i=1}^{l} c_i y_i [\mathbf{w}^T \phi(x_i) + b] + C_u \sum_{j=l+1}^{n} c_j |\mathbf{w}^T \phi(x_j) + b| \right\} \\
& \quad \geq \frac{1}{n} \left\{ C_l \sum_{i=1}^{l} c_i + C_u \sum_{j=l+1}^{n} c_j \right\} - \xi, \quad \forall \mathbf{c} \in \{0, 1\}^n
\end{align*}$$

and any solution $\mathbf{w}^*$ to problem (22) is also a solution to problem (21) (vice versa), with $\xi^* = \frac{C_l}{n} \sum_{i=1}^{l} \xi_i^* + \frac{C_u}{n} \sum_{i=l+1}^{n} \xi_i^*$. 

Changshui Zhang, Bin Zhao
Maximum Margin Embedding

- Traditional embedding methods find the optimal subspace by minimizing some form of average loss or cost.
- MME directly finds the most discriminative subspace, where clusters are most well-separated.
- MME is insensitive to the actual probability distribution of patterns lying further away from the separating hyperplanes.
Maximum Margin Embedding

\[
\min_{\mathbf{y}, \mathbf{w}, b, \xi_i \geq 0} \quad \frac{1}{2} \mathbf{w}^T \mathbf{w} + \frac{C}{n} \sum_{i=1}^{n} \xi_i
\]

s.t. \quad y_i (\mathbf{w}^T \phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \quad i = 1, \ldots, n

\[
A^T \mathbf{w} = 0
\]

\[
\mathbf{y} \in \{-1, +1\}^n
\]

where \( A = [\mathbf{w}^1, \ldots, \mathbf{w}^{r-1}] \) constrains that \( \mathbf{w} \) should be orthogonal to all previously calculated projecting vectors.
Conclusions

Improvements

- No loss in clustering accuracy
- Major improvement on speed
- Handle large real-world datasets efficiently
Conclusions

Future works

- Automatically tune the parameters
- Even larger dataset
References

- Bin Zhao, Fei Wang, Changshui Zhang. Efficient Maximum Margin Clustering Via Cutting Plane Algorithm. SDM 2008
Thanks for Listening