Efficient MultiClass Maximum Margin Clustering

Bin Zhao, Fei Wang, Changshui Zhang

Dept. Automation, Tsinghua Univ.

ICML, July 6, 2008
Helsinki, Finland
Outline

1. Two-Class Maximum Margin Clustering
2. MultiClass Maximum Margin Clustering
3. Theoretical Analysis
4. Experimental Results
5. Conclusions

Bin Zhao, Fei Wang, Changshui Zhang
Efficient MultiClass Maximum Margin Clustering
Support Vector Machine

Given $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, $\mathbf{y} = (y_1, \ldots, y_n) \in \{-1, +1\}^n$, SVM finds a hyperplane $f(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b$ by solving

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

(1)
MMC targets to find not only the optimal hyperplane $(\mathbf{w}^*, b^*)$, but also the optimal labeling vector $\mathbf{y}^*$

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

$$\text{s.t.} \quad y_i(\mathbf{w}^T \phi(x_i) + b) \geq 1 - \xi_i$$
$$\quad \xi_i \geq 0 \quad i = 1, \ldots, n$$
Outline

1. Two-Class Maximum Margin Clustering
2. MultiClass Maximum Margin Clustering
3. Theoretical Analysis
4. Experimental Results
5. Conclusions
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. \quad \forall i = 1, \ldots, n, r = 1, \ldots, k

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

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

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

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

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

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

\[
\sum_{p=1}^{k} w_p^T x_i \prod_{q=1, q \neq p}^{k} I(w_p^T x_i > w_q^T x_i) + \prod_{q=1, q \neq r}^{k} I(w_r^T x_i > w_q^T x_i) - w_r^T x_i \geq 1 - \xi_i
\]

where \( I(\cdot) \) is the indicator function and the label for sample \( x_i \) is determined as \( y_i = \sum_{p=1}^{k} p \prod_{q=1, q \neq p}^{k} I(w_p^T x_i > w_q^T x_i) \)
Theorem

Problem (5) can be equivalently formulated as problem (6), with
\[ \xi^* = \frac{1}{n} \sum_{i=1}^{n} \xi_i^* \cdot \]
\[ \min_{w_1, \ldots, w_k, \xi} \frac{1}{2} \beta \sum_{p=1}^{k} \|w_p\|^2 + \xi \]
\[ \text{s.t. } \forall \mathbf{c}_i \in \{ \mathbf{e}_0, \mathbf{e}_1, \ldots, \mathbf{e}_k \}, \ i = 1, \ldots, n \]
\[ \frac{1}{n} \sum_{i=1}^{n} \left\{ \mathbf{c}_i^T \mathbf{e} \sum_{p=1}^{k} w_p^T x_i z_{ip} + \sum_{p=1}^{k} c_{ip} (z_{ip} - w_p^T x_i) \right\} \geq \frac{1}{n} \sum_{i=1}^{n} \mathbf{c}_i^T \mathbf{e} - \xi \]

where \( z_{ip} = \prod_{q=1, q \neq p}^{k} I(w_p^T x_i > w_q^T x_i) \) and each constraint \( \mathbf{c} \) is represented as a \( k \times n \) matrix \( \mathbf{c} = (\mathbf{c}_1, \ldots, \mathbf{c}_n) \).
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 (6) up to a precision of $\epsilon$. 
Cutting Plane Algorithm [J. E. Kelley 1960, T. Joachims 2006]

- Starts with an empty constraint subset $\Omega$
- Computes the optimal solution to problem (6) subject to the constraints in $\Omega$
- Finds the most violated constraint in problem (6) and adds it into the subset $\Omega$
- Stops when no constraint in (6) is violated by more than $\epsilon$

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

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

Theorem

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

\[
c_i = \begin{cases} 
  e_{r^*} & \text{if } (w_{p^*}^T x_i - w_{r^*}^T x_i) < 1 \\
  0 & \text{otherwise}
\end{cases}, \quad i = 1, \ldots, n
\] (8)
Enforcing the Class Balance Constraint

To avoid trivially “optimal” solutions

\[
\begin{align*}
\min_{w_1, \ldots, w_k, \xi \geq 0} & \quad \frac{1}{2} \beta \sum_{p=1}^{k} ||w_p||^2 + \xi \\
\text{s.t.} & \quad \frac{1}{n} \sum_{i=1}^{n} \left\{ c_i^T e \sum_{p=1}^{k} w_p^T x_i z_{ip} + \sum_{p=1}^{k} c_{ip} (z_{ip} - w_p^T x_i) \right\} \\
& \quad \geq \frac{1}{n} \sum_{i=1}^{n} c_i^T e - \xi, \forall [c_1, \ldots, c_n] \in \Omega \\
& \quad -l \leq \sum_{i=1}^{n} w_p^T x_i - \sum_{i=1}^{n} w_q^T x_i \leq l, \forall p, q = 1, \ldots, k
\end{align*}
\]

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

$$\min_z f_0(z) - g_0(z)$$

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

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

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

$$\begin{align*}
\min_z f_0(z) - T_1\{g_0, z_t\}(z) \\
\text{s.t. } f_i(z) - T_1\{g_i, z_t\}(z) \leq c_i \quad i = 1, \ldots, n
\end{align*}$$

(11)
Optimization via the CCCP

Calculate the subgradients

\[
\partial_{w_r}\left\{ \frac{1}{n} \sum_{i=1}^{n} \left[ \mathbf{c}_i^T \mathbf{e} \sum_{p=1}^{k} \mathbf{w}_p^T \mathbf{x}_i z_{ip} + \sum_{p=1}^{k} c_{ip} z_{ip} \right] \right\} \bigg|_{w = w(t)}
\]

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

By substituting first-order Taylor expansion into problem (9), we obtain a quadratic programming (QP) problem.
Outline

1. Two-Class Maximum Margin Clustering
2. Multi-Class Maximum Margin Clustering
3. Theoretical Analysis
4. Experimental Results
5. Conclusions
Justification of \textit{CPM3C}

\textbf{Theorem}

\textit{For any dataset } $\mathcal{X} = (x_1, \ldots, x_n)$ \textit{and any } $\epsilon > 0$, 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.}
Theorem

Each iteration of CPM3C takes time $O(snk)$ for a constant working set size $|\Omega|$.

Theorem

For any $\epsilon > 0$, $\beta > 0$, and any dataset $X = \{x_1, \ldots, x_n\}$ with samples belonging to two different classes, the CPM3C algorithm terminates after adding at most $\frac{R}{\epsilon^2}$ constraints, where $R$ is a constant number independent of $n$ and $s$. 
The theorem states that for any dataset $\mathcal{X} = \{x_1, \ldots, x_n\}$ with $n$ samples belonging to 2 classes and sparsity of $s$, and any fixed value of $\beta > 0$ and $\epsilon > 0$, the CPM3C algorithm takes time $O(sn)$ to converge.
## Clustering Accuracy Comparison: Two-Class Scenario

<table>
<thead>
<tr>
<th>Data</th>
<th>KM</th>
<th>NC</th>
<th>MMC</th>
<th>GMC</th>
<th>SVR</th>
<th>CPM3C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dig 3-8</td>
<td>94.68</td>
<td>65.00</td>
<td>90.00</td>
<td>94.40</td>
<td>96.64</td>
<td><strong>96.92</strong></td>
</tr>
<tr>
<td>Dig 1-7</td>
<td>94.45</td>
<td>55.00</td>
<td>68.75</td>
<td>97.8</td>
<td>99.45</td>
<td><strong>100.0</strong></td>
</tr>
<tr>
<td>Dig 2-7</td>
<td>96.91</td>
<td>66.00</td>
<td>98.75</td>
<td>99.50</td>
<td><strong>100.0</strong></td>
<td><strong>100.0</strong></td>
</tr>
<tr>
<td>Dig 8-9</td>
<td>90.68</td>
<td>52.00</td>
<td>96.25</td>
<td>84.00</td>
<td>96.33</td>
<td>97.74</td>
</tr>
<tr>
<td>Letter</td>
<td>82.06</td>
<td>76.80</td>
<td>-</td>
<td>-</td>
<td>92.80</td>
<td>94.47</td>
</tr>
<tr>
<td>UCISat</td>
<td>95.93</td>
<td>95.79</td>
<td>-</td>
<td>-</td>
<td>96.82</td>
<td><strong>98.48</strong></td>
</tr>
<tr>
<td>Text-1</td>
<td>50.53</td>
<td>93.79</td>
<td>-</td>
<td>-</td>
<td><strong>96.82</strong></td>
<td>95.00</td>
</tr>
<tr>
<td>Text-2</td>
<td>50.38</td>
<td>91.35</td>
<td>-</td>
<td>-</td>
<td>93.99</td>
<td><strong>96.28</strong></td>
</tr>
<tr>
<td>UCIDig</td>
<td>96.38</td>
<td>97.57</td>
<td>-</td>
<td>-</td>
<td>98.18</td>
<td><strong>99.38</strong></td>
</tr>
<tr>
<td>MNIST¹</td>
<td>89.21</td>
<td>89.92</td>
<td>-</td>
<td>-</td>
<td>92.41</td>
<td>95.71</td>
</tr>
</tbody>
</table>

¹For UCI digits and MNIST datasets, we give a through 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).
Clustering Accuracy Comparison: MultiClass Scenario

<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>
Speed Comparison: Two-Class Scenario

<table>
<thead>
<tr>
<th>Data</th>
<th>KM</th>
<th>GMC</th>
<th>SVR</th>
<th>CPM3C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dig 3-8</td>
<td>0.51</td>
<td>276.16</td>
<td>19.72</td>
<td>1.10</td>
</tr>
<tr>
<td>Dig 1-7</td>
<td>0.54</td>
<td>289.53</td>
<td>20.49</td>
<td>0.95</td>
</tr>
<tr>
<td>Dig 2-7</td>
<td>0.50</td>
<td>304.81</td>
<td>19.69</td>
<td>0.75</td>
</tr>
<tr>
<td>Dig 8-9</td>
<td>0.49</td>
<td>277.26</td>
<td>19.41</td>
<td>0.85</td>
</tr>
<tr>
<td>Letter</td>
<td>0.08</td>
<td>-</td>
<td>2133</td>
<td>0.87</td>
</tr>
<tr>
<td>UCISat</td>
<td>0.19</td>
<td>-</td>
<td>6490</td>
<td>4.54</td>
</tr>
<tr>
<td>Text-1</td>
<td>66.09</td>
<td>-</td>
<td>930.0</td>
<td>19.75</td>
</tr>
<tr>
<td>Text-2</td>
<td>52.32</td>
<td>-</td>
<td>913.8</td>
<td>16.16</td>
</tr>
</tbody>
</table>
## Speed Comparison: MultiClass Scenario

<table>
<thead>
<tr>
<th>Data</th>
<th>KM</th>
<th>CPM3C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dig 0689</td>
<td>34.28</td>
<td>9.66</td>
</tr>
<tr>
<td>Dig 1279</td>
<td>17.78</td>
<td>17.47</td>
</tr>
<tr>
<td>Cora-DS</td>
<td>839.67</td>
<td>35.31</td>
</tr>
<tr>
<td>Cora-HA</td>
<td>204.43</td>
<td>24.35</td>
</tr>
<tr>
<td>Cora-ML</td>
<td>22781</td>
<td>69.04</td>
</tr>
<tr>
<td>Cora-OS</td>
<td>47931</td>
<td>13.98</td>
</tr>
<tr>
<td>Cora-PL</td>
<td>7791.4</td>
<td>165.0</td>
</tr>
<tr>
<td>WK-CL</td>
<td>672.69</td>
<td>9.534</td>
</tr>
<tr>
<td>WK-TX</td>
<td>766.77</td>
<td>10.53</td>
</tr>
<tr>
<td>WK-WT</td>
<td>4135.2</td>
<td>10.67</td>
</tr>
<tr>
<td>WK-WC</td>
<td>1578.2</td>
<td>9.041</td>
</tr>
<tr>
<td>20-news</td>
<td>2387.8</td>
<td>215.6</td>
</tr>
<tr>
<td>RCVI</td>
<td>428770</td>
<td>587.9</td>
</tr>
</tbody>
</table>
Efficient MultiClass Maximum Margin Clustering

Dataset Size $n$ vs. Speed

Cora & 20News

Cora−DS
Cora−HA
Cora−ML
Cora−OS
Cora−PL
20News
O(n)

WebKB & RCVI

WK−CL
WK−HA
WK−WT
WK−WC
RCVI
O(n)
ε vs. Accuracy

Two-Class Maximum Margin Clustering
MultiClass Maximum Margin Clustering
Theoretical Analysis
Experimental Results
Conclusions

Bin Zhao, Fei Wang, Changshui Zhang
Efficient MultiClass Maximum Margin Clustering
Outline

1. Two-Class Maximum Margin Clustering
2. MultiClass Maximum Margin Clustering
3. Theoretical Analysis
4. Experimental Results
5. Conclusions
Conclusions

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

- **Future works**
  - Automatically tune the parameters
  - Even larger dataset
Thanks for Listening