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

Outline

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

Support Vector Machine

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

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

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 $\mathbf{y}^*$

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

Outline

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

Multi-Class Support Vector Machine [Crammer & Singer 2001]

Given a point set $\mathcal{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ and their labels $\mathbf{y} = (y_1, \ldots, y_n) \in \{1, \ldots, k\}^n$, *SVM* defines a weight vector $\mathbf{w}_p$ for each class $p \in \{1, \ldots, k\}$ and classifies sample $\mathbf{x}$ by $y^* = \arg\max_{y \in \{1,\ldots,k\}} \mathbf{w}_y^T \mathbf{x}$ with the weight vectors obtained as

$$\begin{aligned} \min_{\mathbf{w}_1,\ldots,\mathbf{w}_k,\xi} \quad & \frac{1}{2}\beta \sum_{p=1}^{k} ||\mathbf{w}_p||^2 + \sum_{i=1}^{n} \xi_i \\ s.t. \quad & \forall i = 1, \ldots, n, r = 1, \ldots, k \\ & \mathbf{w}_{y_i}^T \mathbf{x}_i + \delta_{y_i,r} - \mathbf{w}_r^T \mathbf{x}_i \geq 1 - \xi_i \end{aligned} \tag{3}$$

Multi-Class Maximum Margin Clustering

Similar with the binary clustering scenario

$$\begin{aligned} \min_{\mathbf{y}} \min_{\mathbf{w}_1,\ldots,\mathbf{w}_k,\xi} \quad & \frac{1}{2}\beta\sum_{p=1}^{k}||\mathbf{w}_p||^2+\frac{1}{n}\sum_{i=1}^{n}\xi_i \\ s.t. \quad & \forall i=1,\ldots,n, r=1,\ldots,k \\ & \mathbf{w}_{y_i}^T\mathbf{x}_i+\delta_{y_i,r}-\mathbf{w}_r^T\mathbf{x}_i\geq 1-\xi_i \end{aligned} \tag{4}$$

Problem Reformulation I

Theorem

$$\begin{aligned} \min_{\mathbf{w}_1,\ldots,\mathbf{w}_k,\xi} \quad & \frac{1}{2}\beta\sum_{p=1}^{k}||\mathbf{w}_p||^2+\frac{1}{n}\sum_{i=1}^{n}\xi_i \\ s.t. \quad & \forall i=1,\ldots,n, r=1,\ldots,k \\ & \sum_{p=1}^{k}\mathbf{w}_p^T\mathbf{x}_i\prod_{q=1,q\neq p}^{k} I_{(\mathbf{w}_p^T\mathbf{x}_i>\mathbf{w}_q^T\mathbf{x}_i)}+\prod_{q=1,q\neq r}^{k} I_{(\mathbf{w}_r^T\mathbf{x}_i>\mathbf{w}_q^T\mathbf{x}_i)}-\mathbf{w}_r^T\mathbf{x}_i\geq 1-\xi_i \end{aligned} \tag{5}$$

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 Reformulation II

Theorem

Problem (5) can be equivalently formulated as problem (6), with $\xi^* = \frac{1}{n}\sum_{i=1}^{n}\xi_i^*$.

$$\min_{\mathbf{w}_1,\ldots,\mathbf{w}_k,\xi} \frac{1}{2}\beta\sum_{p=1}^{k}||\mathbf{w}_p||^2+\xi \tag{6}$$

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

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

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

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 ϵ.

Cutting Plane Algorithm [J. E. Kelley 1960, T. Joachims 2006]

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

$$\forall \mathbf{c}_i \in \{\mathbf{e}_0, \mathbf{e}_1, \ldots, \mathbf{e}_k\}^n,\ i = 1, \ldots, n \tag{7}$$

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

The Most Violated Constraint

Theorem

Define $p^* = \arg\max_p(\mathbf{w}_p^T\mathbf{x}_i)$ *and* $r^* = \arg\max_{r \neq p^*}(\mathbf{w}_r^T\mathbf{x}_i)$ *for* $i = 1, \ldots, n$*, the most violated constraint could be calculated as follows*

$$\mathbf{c}_i = \begin{cases} \mathbf{e}_{r^*} & \text{if } (\mathbf{w}_{p^*}^T\mathbf{x}_i - \mathbf{w}_{r^*}^T\mathbf{x}_i) < 1 \\ \mathbf{0} & \text{otherwise} \end{cases}, \quad i = 1, \ldots, n \tag{8}$$

Enforcing the Class Balance Constraint

To avoid trivially “optimal” solutions

$$\min_{\mathbf{w}_1,\ldots,\mathbf{w}_k,\xi\geq 0} \quad \frac{1}{2}\beta\sum_{p=1}^{k}||\mathbf{w}_p||^2+\xi \tag{9}$$

$$\begin{aligned}
s.t. \quad & \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbf{c}_i^T\mathbf{e}\sum_{p=1}^{k}\mathbf{w}_p^T\mathbf{x}_i z_{ip}+\sum_{p=1}^{k}c_{ip}(z_{ip}-\mathbf{w}_p^T\mathbf{x}_i)\right\} \\
& \geq \frac{1}{n}\sum_{i=1}^{n}\mathbf{c}_i^T\mathbf{e}-\xi,\ \forall[\mathbf{c}_1,\ldots,\mathbf{c}_n]\in\Omega \\
& -l\leq\sum_{i=1}^{n}\mathbf{w}_p^T\mathbf{x}_i-\sum_{i=1}^{n}\mathbf{w}_q^T\mathbf{x}_i\leq l,\ \forall p,q=1,\ldots,k
\end{aligned}$$

The Constrained Concave-Convex Procedure [A. J. Smola et.al. 2005]

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

$$\begin{array}{ll} \min_{\mathbf{z}} & f_0(\mathbf{z}) - g_0(\mathbf{z}) \\ s.t. & f_i(\mathbf{z}) - g_i(\mathbf{z}) \leq c_i \quad i = 1, \ldots, n \end{array} \tag{10}$$

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 $\mathbf{z}_0$, the *CCCP* computes $\mathbf{z}_{t+1}$ from $\mathbf{z}_t$ by replacing $g_i(\mathbf{z})$ with its first-order Taylor expansion at $\mathbf{z}_t$

$$\begin{aligned} \min_{\mathbf{z}} \ & f_0(\mathbf{z}) - T_1\{g_0, \mathbf{z}_t\}(\mathbf{z}) \\ s.t. \ & f_i(\mathbf{z}) - T_1\{g_i, \mathbf{z}_t\}(\mathbf{z}) \leq c_i \quad i = 1, \ldots, n \end{aligned} \tag{11}$$

Optimization via the *CCCP*

Calculate the subgradients

$$\partial_{\mathbf{w}_r}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{c}_i^T\mathbf{e}\sum_{p=1}^{k}\mathbf{w}_p^T\mathbf{x}_i z_{ip}+\sum_{p=1}^{k}c_{ip}z_{ip}\right]\right\}\Bigg|_{\mathbf{w}=\mathbf{w}^{(t)}} \tag{12}$$

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

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

Outline

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

Justification of *CPM3C*

Theorem

For any dataset $\mathcal{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ *and any* $\epsilon > 0$*, the CPM3C algorithm returns a point* $(\mathbf{w}_1, \ldots, \mathbf{w}_k, \xi)$ *for which* $(\mathbf{w}_1, \ldots, \mathbf{w}_k, \xi + \epsilon)$ *is feasible.*

Time Complexity Analysis

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 $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ 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.

Time Complexity Analysis

Theorem

For any dataset $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ 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.

Outline

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

Clustering Accuracy Comparison: Two-Class Scenario

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

[1] For 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

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

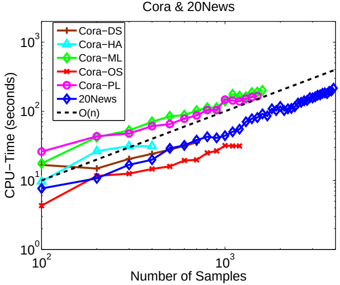
Speed Comparison: Two-Class Scenario

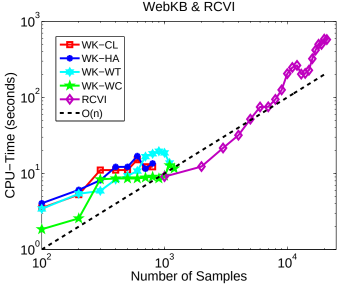
Data	KM	GMC	SVR	CPM3C
Dig 3-8	0.51	276.16	19.72	1.10
Dig 1-7	0.54	289.53	20.49	0.95
Dig 2-7	0.50	304.81	19.69	0.75
Dig 8-9	0.49	277.26	19.41	0.85
Letter	0.08	-	2133	0.87
UCISat	0.19	-	6490	4.54
Text-1	66.09	-	930.0	19.75
Text-2	52.32	-	913.8	16.16

Speed Comparison: MultiClass Scenario

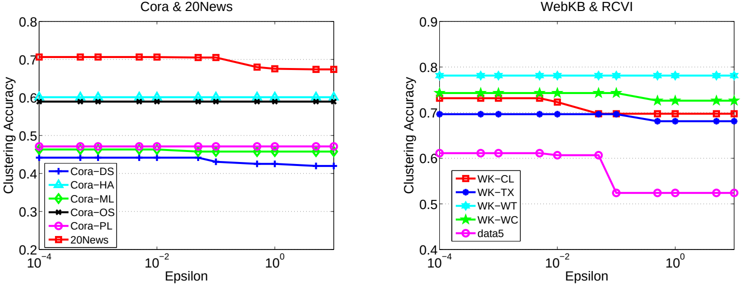
Data	KM	CPM3C
Dig 0689	34.28	9.66
Dig 1279	17.78	17.47
Cora-DS	839.67	35.31
Cora-HA	204.43	24.35
Cora-ML	22781	69.04
Cora-OS	47931	13.98
Cora-PL	7791.4	165.0
WK-CL	672.69	9.534
WK-TX	766.77	10.53
WK-WT	4135.2	10.67
WK-WC	1578.2	9.041
20-news	2387.8	215.6
RCVI	428770	587.9

Dataset Size n vs. Speed

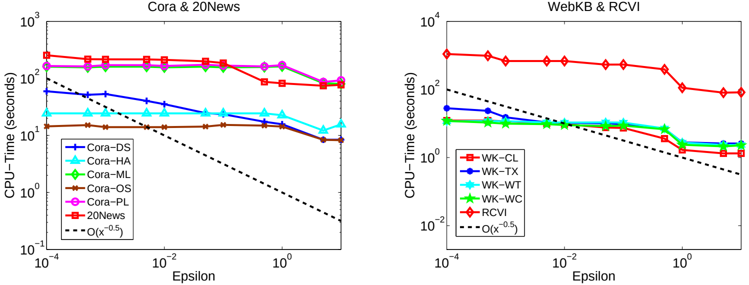




ϵ vs. Accuracy



ϵ vs. Speed



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