

Multiclass classification by L_1 norm Support Vector Machine

Sandor Szedmak,
joint work with
John Shawe-Taylor, Craig J. Saunders and David R. Hardoon

ISIS Group
School of Electronics and Computer Science
University of Southampton

PASCAL Project

June 14, 2004

Contents

- A (short) introduction
- Minimax framework on One-Vs-All
- How to solve it?
 - L_1 norm SVM
 - Minimax formulation
 - The framework of the solution
 - The “trick”

A (short) introduction

Multiclass classification:

- There are several approach, but no good “idea”, theoretical foundation
- Approaches:
 - Separated subproblems, e.g. One-Vs-All, All-Vs-All,
 - Model based solution for the full problem, e.g. CCA(KCCA),
- Difficulties on large scale problems($\gg 1000$ cases, $\gg 1000$ features)

Our objective to find a solution framework for the full problem with enormous size.

Minimax framework on One-Vs-All

- Consider the subproblems of a One-Vs-All classification
- Maximise the minimum margin on all of the subproblems

The minimax framework gives a clear structure and it recalls several well founded ideas of the optimisation theory.

One can apply the schema of the formulation for solving special tasks. We will present an example for clustering.

How to solve it?

It seems, this framework can not be a tractable optimisation problem!

- If the subproblems are linear, L_1 or L_∞ norm based, then we can exploit efficiently the potential sparsity of the formulation
- We can apply decomposition method to cut down the complexity,

We show a solution schema which is able to overcome on the difficulties caused by the enormous size.

The algorithm is presented for the L_1 norm SVM case.

Remark 1. *The solution scheme is applicable for L_∞ norm SVM or Linear Programming Boosting as well.*

L_1 norm SVM

$$\begin{aligned} \min_{w_+, w_-, b, \xi, \eta} \quad & e^T w_+ + e^T w_- + D e^T \xi + D e^T \eta \\ \text{subject to} \quad & \\ & X_+ w_+ - X_+ w_- + b \geq e - \xi, \\ & -X_- w_+ + X_- w_- - b \geq e - \eta, \\ & \xi \geq 0, \eta \geq 0, \\ & w_+ \geq 0, w_- \geq 0. \end{aligned} \tag{1}$$

$$w = w_+ - w_-, \quad \|w\|_1 = e^T w_+ + e^T w_-. \tag{2}$$

Minimax formulation

$$\min_{\{\nu, w_k^+, w_k^-, b_k, \xi_k, \eta_k\}} \quad \nu + \sum_k^K (D_k e^T \xi_k + D_k e^T \eta_k)$$

subject to

$X_1 w_1^+ - X_1 w_1^- + b_1 + \xi_1,$		$\geq e,$
$-\bar{X}_1 w_1^+ + \bar{X}_1 w_1^- - b_1 + \eta_1,$		$\geq e,$
\dots		\vdots
	$X_K w_K^+ + X_K w_K^- + b_K + \xi_K,$	$\geq e,$
	$-\bar{X}_K w_K^+ + \bar{X}_K w_K^- - b_K + \eta_K,$	$\geq e,$
$e^T w_1^+ + e^T w_1^-$		$\leq \nu$
\dots		$\leq \nu$
	$e^T w_K^+ + e^T w_K^-$	$\leq \nu$
$\xi_1 \geq 0, \eta_1 \geq 0,$	\dots	$\xi_K \geq 0, \eta_K \geq 0,$
$w_1^+ \geq 0, w_1^- \geq 0,$	\dots	$w_K^+ \geq 0, w_K^- \geq 0,$

(3)

Dual of the Minimax

$$\begin{array}{rcl}
 \max_{\{\alpha_k, \beta_k, \gamma_k\}} & & \sum_k^K (e^T \alpha_k + e^T \beta_k) \\
 & \text{subject to} & \\
 \hline
 \| + X_1^T \alpha_1 - \bar{X}_1^T \beta_1 \|_\infty & & \leq \gamma_1 \\
 \dots & & \vdots \\
 \| + X_K^T \alpha_K - \bar{X}_K^T \beta_K \|_\infty & & \leq \gamma_K \\
 \hline
 e^T \alpha_1 - e^T \beta_1 & & = 0 \\
 \dots & & \vdots \\
 e^T \alpha_K - e^T \beta_K & & = 0 \\
 \hline
 0 \leq \alpha_1 \leq D_1, 0 \leq \beta_1 \leq D_1 & \dots & 0 \leq \alpha_K \leq D_K, 0 \leq \beta_K \leq D_K \\
 \hline
 \gamma_1 + & \dots & + \gamma_K & = 1 \\
 \hline
 \gamma_1 \geq 0, & \dots & \gamma_K \geq 0, & \\
 & & & (4)
 \end{array}$$

Bennett et al. (2000) [2], Mangasarian (1999) [4]

A clustering example

The content of the formal problem can be changed by giving special meaning to the matrices $\{(X_1, \bar{X}_1), \dots, (X_K, \bar{X}_K)\}$.

- Given a set of images $\{A_1, \dots, A_K\}$ by sets of feature vectors as prototypes of some classes.
- We received an image B and the question what are the prototypes (or subset of the prototypes) most similar or dissimilar to B .
- The prototypes and the test image have different number of feature vectors but the number of the components are the same (e.g. interest points + SHIFT features).
- Set up pairs such that $(B, A_1), \dots, (B, A_K)$ and solve the dual problem

This type of problems gives a special equilibrium solution where all effects are considered simultaneously.

Find less dissimilar prototypes

$$\begin{array}{rcc}
 \max_{\{\alpha_k, \beta_k, \gamma_k\}} & & \sum_k^K (e^T \alpha_k + e^T \beta_k) \\
 & \text{subject to} & \\
 \hline
 \| + B^T \alpha_1 - A_1^T \beta_1 \|_\infty & & \leq \gamma_1 \\
 \dots & & \vdots \\
 \| + B^T \alpha_K - A_K^T \beta_K \|_\infty & & \leq \gamma_K \\
 \hline
 e^T \alpha_1 - e^T \beta_1 & & = 0 \\
 \dots & & \vdots \\
 e^T \alpha_K - e^T \beta_K & & = 0 \\
 \hline
 0 \leq \alpha_1 \leq D_1, 0 \leq \beta_1 \leq D_1 & \dots & 0 \leq \alpha_K \leq D_K, 0 \leq \beta_K \leq D_K \\
 \hline
 \gamma_1 + & \dots & + \gamma_K & = 1 \\
 \hline
 \gamma_1 \geq 0, & \dots & \gamma_K \geq 0, & \\
 & & & (5)
 \end{array}$$

Master problem

$$\min_{\{\alpha_k, \beta_k, \gamma_k\}} \quad - \sum_k^K (e^T \alpha_k + e^T \beta_k)$$

subject to

$+X_1^T \alpha_1 - \bar{X}_1^T \beta_1$		$\leq \gamma_1,$	(6)
$-X_1^T \alpha_1 + \bar{X}_1^T \beta_1$		$\leq \gamma_1,$	
\dots		\vdots	
	$+X_K^T \alpha_K - \bar{X}_K^T \beta_K$	$\leq \gamma_K,$	
	$-X_K^T \alpha_K + \bar{X}_K^T \beta_K$	$\leq \gamma_K,$	
$\alpha_1 \geq 0, \beta_1 \geq 0, \gamma_1 \geq 0, \dots \alpha_K \geq 0, \beta_K \geq 0, \gamma_K \geq 0,$			

Subproblem

$$\min_{\gamma_k, \alpha_k, \beta_k} \quad (c - \pi A)^T (\alpha_1, \dots, \alpha_K, \beta_1, \dots, \beta_K, \gamma_1, \dots, \gamma_K)$$

subject to

$$\begin{array}{rcl}
 e^T \alpha_1 - e^T \beta_1 & & = 0 \\
 \dots & & \vdots \\
 & e^T \alpha_K - e^T \beta_K & = 0 \\
 \hline
 0 \leq \alpha_1 \leq D_1, 0 \leq \beta_1 \leq D_1 \ \dots \ 0 \leq \alpha_K \leq D_K, 0 \leq \beta_K \leq D_K & & \\
 \hline
 \gamma_1 + & \dots & + \gamma_K \quad = 1 \\
 \hline
 \gamma_1 \geq 0, & \dots & \gamma_K \geq 0,
 \end{array} \tag{7}$$

The components of the objective function in the subproblem

$$c^T = (\underbrace{-e^T, \dots, -e^T}_{1, \dots, K}, \underbrace{-e^T, \dots, -e^T}_{1, \dots, K}, \underbrace{0, \dots, 0}_{1, \dots, K}),$$

$$A = \begin{bmatrix} -X_1 & & +\bar{X}_1 & & +e \\ +X_1 & & -\bar{X}_1 & & +e \\ & \dots & & \dots & \\ & & -X_K & & +\bar{X}_1 & +e \\ & & +X_K & & -\bar{X}_1 & +e \end{bmatrix} \quad (8)$$

and π is the current dual solution of the master in the iteration.

Solution of the subproblem 1

If there is an efficient solution for the subproblem then the entire optimisation task becomes tractable.

We have two types of critical, time consuming tasks:

1. to compute πA in the objective function and Ax in violation,
2. to find the optimum solution knowing the vector $(c - \pi A)$.

The answers:

1. exploiting the diagonal block structure of A πA and Ax can be computed efficiently,
2. there is an $O(n \log(n))$ algorithm which gives the optimum solution, where n is the number of variables.

Solution of the subproblem 2

$$\hat{c} = (c - \pi A) = (\hat{c}_\alpha, \hat{c}_\beta, \hat{c}_\gamma)$$

The problems that we have to solve:

$\begin{aligned} & \min_{\alpha_k, \beta_k} \hat{c}_{\alpha_k}^T \alpha_k + \hat{c}_{\beta_k}^T \beta_k \\ & \text{subject to} \\ & e^T \alpha_k - e^T \beta_k = 0, \\ & 0 \leq \alpha_k \leq D_k, \\ & 0 \leq \beta_k \leq D_k, \\ & k = 1, \dots, K, \end{aligned}$	$\begin{aligned} & \min_{\gamma} \hat{c}_\gamma^T \gamma \\ & \text{subject to} \\ & e^T \gamma = 1, \gamma \geq 0. \end{aligned}$
---	--

(10)

The index vectors I, J give ascending order to \hat{c}_α and \hat{c}_β

$$(\hat{c}_\alpha)_{I_i} \leq (\hat{c}_\alpha)_{I_j}, (\hat{c}_\beta)_{J_i} \leq (\hat{c}_\beta)_{J_j}, \text{ if } i < j \quad (11)$$

Let $T = 1, \dots, \min(|I|, |J|)$ be an index such that

$$(\hat{c}_\alpha)_{I_t} + (\hat{c}_\beta)_{J_t} \begin{cases} < 0 & \text{if } i \leq T, \\ \geq 0 & \text{if } i > T. \end{cases} \quad (12)$$

Solution of the subproblem 3

Then for every $t = 1, \dots, T$ in the optimum of the subproblem the following holds

$$\begin{aligned} (\alpha_k)_{I_t} &= D_k, (\beta_k)_{J_t} = D_k & t \leq T \\ (\alpha_k)_{I_t} &= 0, (\beta_k)_{J_t} = 0 & t > T. \end{aligned} \tag{13}$$

The optimum solution for γ

$$\gamma_k = \begin{cases} 1 & \text{if } k = \arg \min_k (\hat{c}_\gamma)_k, \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

Experiments

	Calttech data	Columbia 100
Number of images	3691	7200
Average number of features	2112	144
Number of categories	5	100
Size of the master problem	21125×9230	28900×360100
Solution time(average)	1650s	3390s

Implementation

Our algorithm is based on the Open Source implementation of the Volume Algorithm. The sources with several other optimisation packages are available on the web site:

<http://www.coin-or.org>

<http://oss.software.ibm.com/developerworks/opensource/coin/index.html>

A “tailored” version with Matlab interface is available at UOS.

Bibliography

- [1] F. Barahona and R. Anbil. The volume algorithm: producing primal solutions with a subgradient method. In *IBM Research Report*. IBM, 1998. http://www.research.ibm.com/resources/paper_search.shtml.
- [2] K.P. Bennett and E. Breidensteiner. Duality and geometry in svm classifiers. In *Proceedings of the Seventeenth International Conference on Machine Learning, Pat Langley Editor*, pages 57–64. Morgan Kaufmann, San Francisco, 2000.
- [3] G. B. Dantzig and P. Wolfe. Decomposition principle for linear programs. *Operations Research*, 8, 1960.
- [4] O.L. Mangasarian. Arbitrary-norm separating plane. *Operations Research Letters*, 24,1:15–23, 1999.