Towards Optimal Discriminating Order for Multiclass Classification

Dong Liu, Shuicheng Yan, Yadong Mu, Xian-Sheng Hua,

Shih-Fu Chang and Hong-Jiang Zhang

Harbin Institute of Technology , China
National University of Singapore, Singapore
Microsoft Research Asia, China
Columbia University, USA

Outline

- Introduction
- Our work
- Experiments
- Conclusion and Future work

Introduction

Multiclass Classification

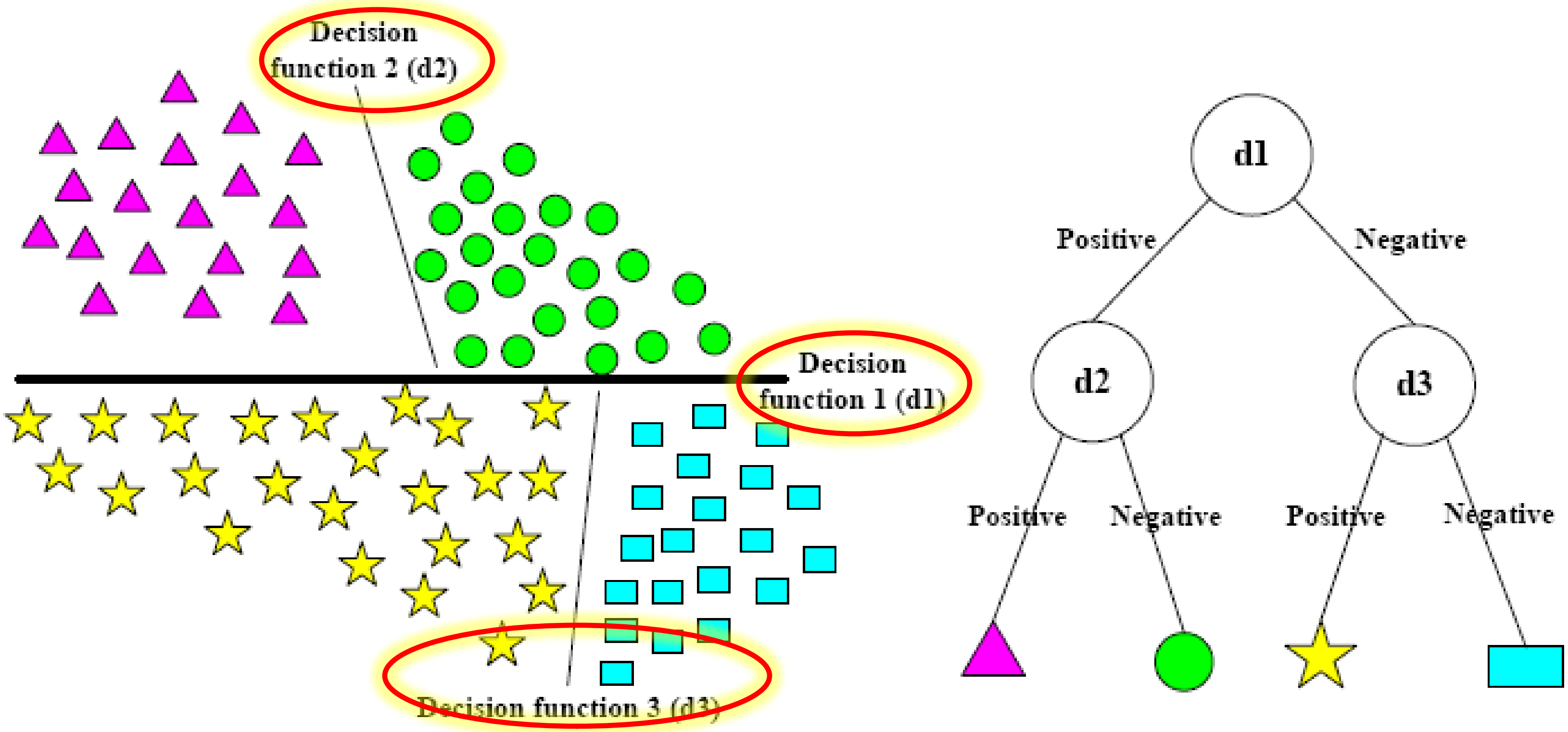
- Supervised multiclass learning problem
  - Accurately assign class labels to instances, where the label set contains at least <u>three</u> elements.
- Important in various applications
  - Natural Language processing, computer vision, computational biology.

Multiclass Classification (con't)

- Discriminate samples from N (N>2) classes.
- Implemented in a stepwise manner:
  - A <u>subset</u> of the N classes are discriminated at first.
  - Further discrimination of the <u>remaining</u> classes.
  - Until <u>all</u> classes can be discriminated.

Multiclass Discriminating Order

- An *approximate discriminating order* is <u>critical</u> for multiclass classification, esp. for linear classifiers.
- E.g., the 4-class data CANNOT be well separated unless using the discriminating order shown here.



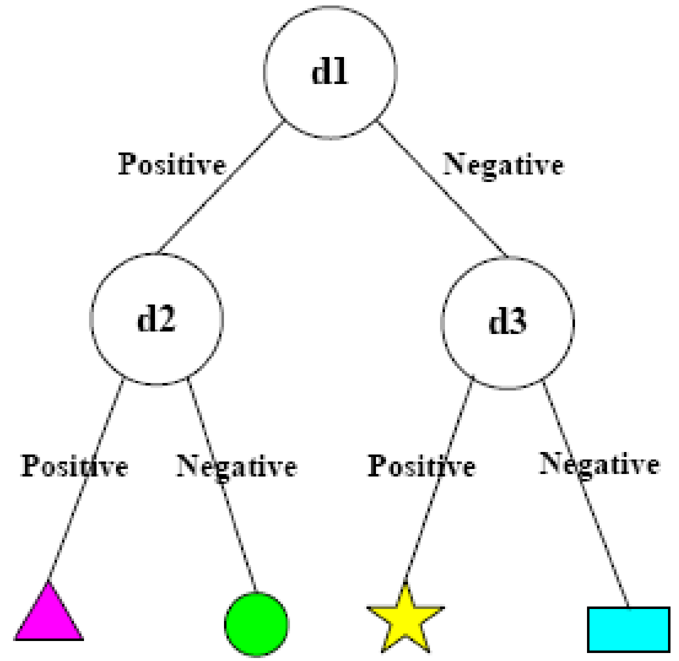
Many Multiclass Algorithms

- One-Vs-All SVM (OVA SVM)
- One-Vs-One SVM (OVO SVM)
- DAGSVM
- Multiclass SVM in an all-together optimization formulation
- Hierarchical SVM
- Error-Correcting Output Codes
- ……

These existing algorithms DO NOT take the discriminating order into consideration, which directly motivates our work here.

Sequential Discriminating Tree

- Derive the optimal discriminating order through a *hierarchical binary partitioning* of the classes.
  - Recursively partition the data such that samples in the same class are grouped into the same subset.
- Use a *binary tree* architecture to represent the discriminating order:
  - Root node: the first discriminating function.
  - Leaf node: final decision of one specific class.



Sequential Discriminating Tree (SDT)

Tree Induction

- Key ingredient : how to perform binary partition at each non-leaf node.
  - Training samples in the same class should be <u>grouped together</u>.
  - The partition function should have a <u>large margin</u> to ensure the generalization ability.
- We employ a <u>constrained large margin binary clustering algorithm</u> as the binary partition procedure at each node of SDT.

Our Work

Constrained Clustering

- Notations

- A collection of samples $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{n}$
- Binary partition hyperplane $f(\mathbf{x}_i) = \omega^\top \mathbf{x}_i + b$
- Constraint set Θ_s
- A constraint indicating that two training samples (i and j) are from the same class $(i, j) \in \Theta_s$
- which side of the hyperplane x_{i} locates y_i

$y_i = +1$ indicates that $\mathbf{x}_i$ is at the positive side

$y_i = -1$ shows that $\mathbf{x}_i$ is at the negative side.

Our Work

Constrained Clustering (con't)

- Objective function

$$\mathcal{J}_\omega = \Omega(\omega) + \lambda_1 \sum_i \ell\left(-y_i f(\mathbf{x}_i)\right) + \lambda_2 \sum_{(i,j)\in\Theta_s} \hbar((i,j)),$$

- Regularization term: $\Omega(\omega) = \frac{1}{2}\|\omega\|^2$
- Hinge loss term: $\ell(x) = (1-x)_+$

Enforce a large margin between samples of different classes.

- Constraint loss term: $\hbar((i,j)) = \begin{cases} 0, & y_i = y_j, \\ (-y_i y_j)_+, & y_i \neq y_j. \end{cases}$

Enforce samples of the same class to be partitioned into the same side of the hyperplane.

Our Work

Constrained Clustering (con't)

- Objective Function

$$
\begin{aligned}
\min_{\omega,b,\xi,\zeta,y} \quad & \frac{1}{2}\|w\|^2 + \frac{\lambda_1}{n}\sum_i \xi_i + \frac{\lambda_2}{n}\sum_{(i,j)\in\Theta_s} \zeta_{ij} \\
s.t. \quad & y_i(\omega^T \mathbf{x}_i + b) + \xi_i \geq 1,\ \xi_i \geq 0,\ \forall\ i, \\
& y_i y_j + \zeta_{ij} \geq 0,\ \zeta_{ij} \geq 0,\ \forall (i,j) \in \Theta_s.
\end{aligned}
$$

- Kernelization

$$
\begin{aligned}
\min_{\alpha,b,\xi,\zeta} \quad & \frac{1}{2}\alpha^T G \alpha + \frac{\lambda_1}{n}\sum_i \xi_i + \frac{\lambda_2}{n}\sum_{(i,j)\in\Theta_s} \zeta_{ij} && (4) \\
s.t. \quad & |\alpha^T k_i + b| + \xi_i \geq 1,\ \forall\ i, && (5) \\
& (\alpha^T k_i + b)(\alpha^T k_j + b) + \zeta_{ij} \geq 0, && (6) \\
& \xi_i \geq 0,\ \forall\ i, \\
& \zeta_{ij} \geq 0,\ \forall\ (i,j) \in \Theta_s,
\end{aligned}
$$

Our Work

Optimization

$$\min_{\alpha,b,\xi,\zeta} \quad \frac{1}{2}\alpha^T G\alpha + \frac{\lambda_1}{n}\sum_i \xi_i + \frac{\lambda_2}{n}\sum_{(i,j)\in\Theta_s}\zeta_{ij} \qquad (4)$$

$$s.t. \quad |\alpha^T k_i + b| + \xi_i \geq 1,\ \forall\ i, \qquad (5)$$

$$(\alpha^T k_i + b)(\alpha^T k_j + b) + \zeta_{ij} \geq 0, \qquad (6)$$

$$\xi_i \geq 0,\ \forall\ i,$$

$$\zeta_{ij} \geq 0,\ \forall\ (i,j) \in \Theta_s,$$

- Optimization Procedure
  - (4) is convex, (5) and (6) can be expressed as the difference of two convex functions.
  - Can be solved with Constrained Concave-Convex Procedure (CCCP).

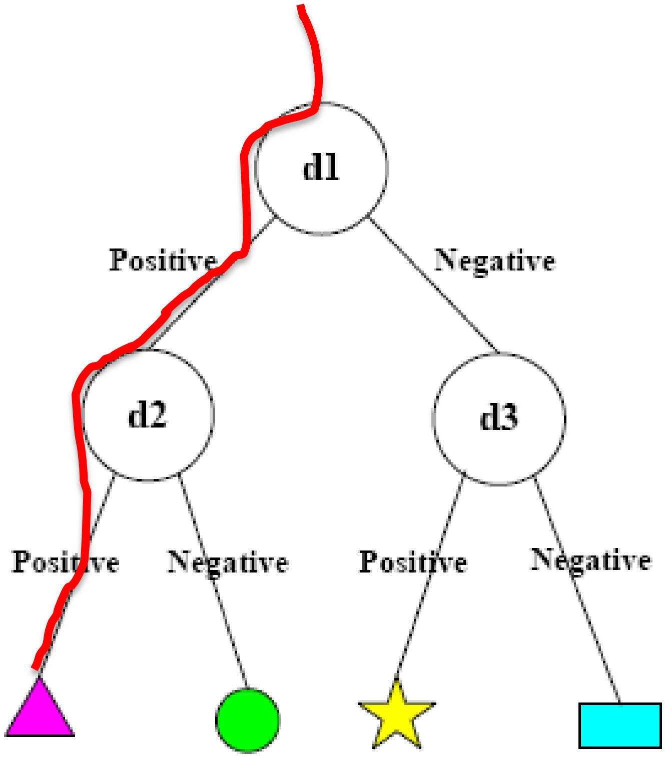
The induction of SDT

- Input: N-class training data T.
- Output: SDT.
  - Partition T into two non-overlapping subsets P and Q using the large margin binary partition procedure.
  - Repeat partitioning subsets P and Q respectively until all obtained subsets only contain training samples from a single class.

Our Work

Prediction

- Evaluate the binary discriminating function at each node of SDT.
- A node is exited via the left edge if the value of the discriminating function is non-negative.
- Or the right edge if the value is negative.



Our Work

Algorithmic Analysis

- Time Complexity

$$T_{SDT} \le \sum_{i=0}^{\lfloor \log_2(N) - 1 \rfloor + 1} (\beta n) = (\lfloor \log_2(N) - 1 \rfloor + 2)\beta n.$$

proportionality constant : β Training set size : n

- Error Bound of SDT

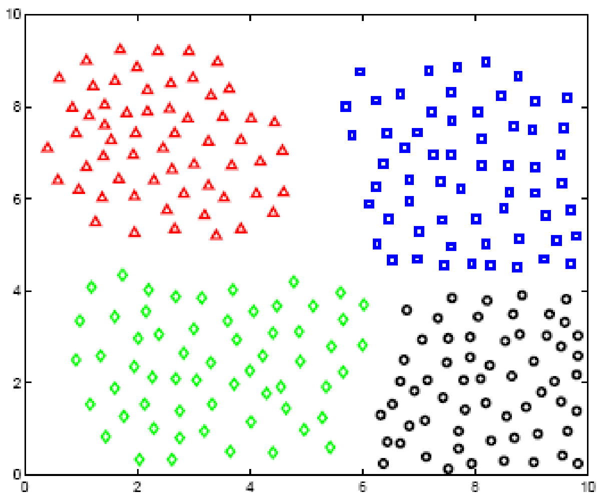
Theorem 3. *Suppose we are able to classify a random n sample of labeled examples using a directed acyclic graph on N classes containing K decision nodes with margins γ_i at node i, then we can bound the generalization error with probability greater than $1 - \delta$ to be less than*

$$\frac{130R^2}{n}(D' log(4en) log(4n) + log\frac{2(2n)^{N-1}}{\delta}),$$

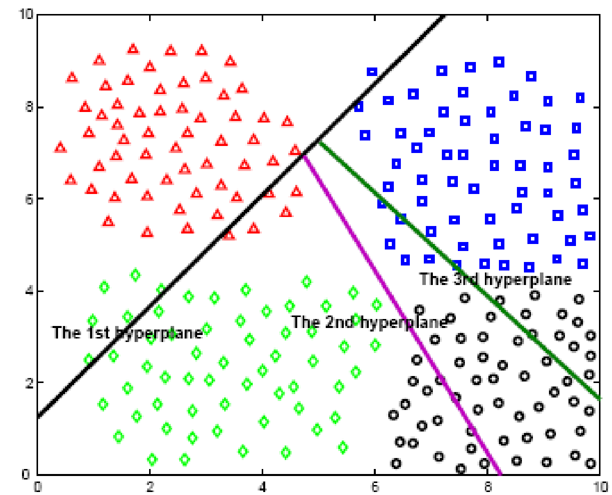
where $D' = \sum_{i=1}^{K} \frac{1}{\gamma_i^2}$, e is the Napierian base, and R is the radius of a ball containing the support of the distribution.

Experiments

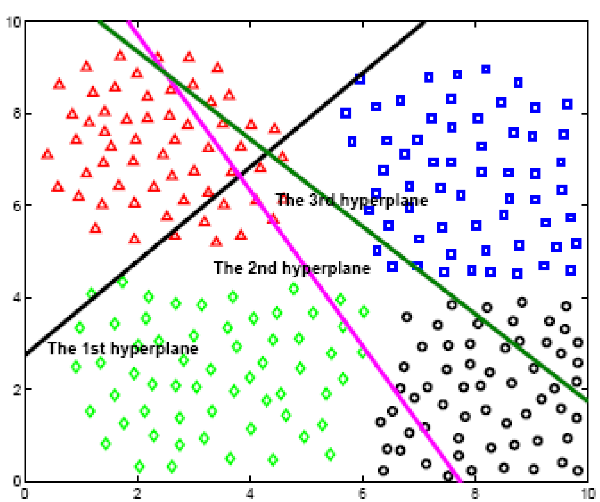
Exp-I: Toy Example



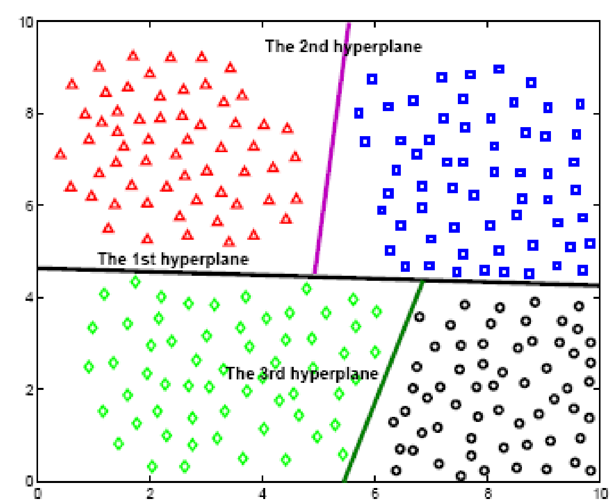
(a) 4-class Toy Data



(b) Hierarchical SVM



(c) OVA SVM



(d) SDT

Exp-II: Benchmark Tasks

- 6 benchmark UCI datasets
  - With pre-defined training/testing splits
  - Frequently used for multiclass classification

Dataset	#training/testing data	#class	#dim.
iris	150/0	3	4
glass	214/0	6	13
vowel	528/0	11	10
vehicle	846/0	4	18
segment	2310/0	7	19
satimage	4435/2000	6	36

Exp-II: Benchmark Tasks (con't)

- In terms of classification accuracy
  - Linear vs. RBF kernel.

Linear Kernel	OVA SVM	OVO SVM	DAGSVM	C&S SVM	Hierarchical SVM	SDT
iris	96.00	97.33	96.67	96.67	97.33	**98.00**
glass	60.28	66.82	61.23	64.95	65.32	**68.49**
vowel	50.95	80.49	81.03	82.57	81.01	**83.75**
vehicle	78.72	81.09	80.13	78.72	79.82	**82.83**
segment	92.47	95.24	94.38	95.37	93.83	**97.75**
satimage	80.35	85.50	86.30	85.15	**87.15**	86.20

RBF Kernel	OVA SVM	OVO SVM	DAGSVM	C&S SVM	Hierarchical SVM	SDT
iris	96.67	97.33	96.67	96.67	97.33	**98.00**
glass	71.76	71.47	72.96	70.87	72.61	**73.16**
vowel	97.79	98.93	98.26	98.85	98.18	**97.02**
vehicle	86.63	86.64	86.32	87.12	86.13	**87.26**
segment	96.78	97.13	97.24	96.88	97.16	**97.53**
satimage	91.45	91.30	91.25	92.35	92.10	**92.45**

Exp-III: Image Categorization

- In terms of classification accuracy and standard derivation
  - COREL image dataset (2,500 images, 255-dim color feature).
  - Linear vs. RBF kernel.

Linear kernel	*accuracy*	RBF kernel	*accuracy*
OVA SVM	66.79 ± 2.13	OVA SVM	70.12 ± 3.31
OVO SVM	71.17 ± 2.25	OVO SVM	75.81 ± 3.62
DAGSVM	69.09 ± 2.74	DAGSVM	75.55 ± 3.63
C&S	68.59 ± 2.16	C&S	73.86 ± 3.03
HierSVM	70.12 ± 2.37	HierSVM	72.27 ± 2.96
SDT	$\mathbf{73.26 \pm 1.98}$	SDT	$\mathbf{77.25 \pm 3.09}$

Exp-IV: Text Categorization

- In terms of classification accuracy and standard derivation
  - 20 Newsgroup dataset (2,000 documents, 62, 061 dim tf-idf feature).
  - Linear vs. RBF kernel.

Linear Kernel	*accuracy*	RBF Kernel	*accuracy*
OVA SVM	51.93 ± 5.72	OVA SVM	52.83 ± 5.93
OVO SVM	57.23 ± 6.82	OVO SVM	60.05 ± 2.74
DAGSVM	59.00 ± 6.79	DAGSVM	67.67 ± 3.67
C&S	55.34 ± 6.26	C&S	66.75 ± 2.96
HierSVM	61.71 ± 5.51	HierSVM	68.26 ± 2.43
SDT	63.23 ± 5.27	SDT	68.72 ± 3.04

Conclusions

- Sequential Discriminating Tree (SDT)
  - Towards the optimal discriminating order for multiclass classification.
  - Employ the constrained large margin clustering algorithm to infer the tree structure.
  - Outperform the state-of-the-art multiclass classification algorithms.

Future work

- Seeking the optimal learning order for
  - Unsupervised clustering
  - Multiclass Active Learning
  - Multiple Kernel Learning
  - Distance Metric Learning
  - ....

Question?

dongliu.hit@gmail.com