Hierarchical Document Categorization with Support Vector Machine

Lijuan Cai, Thomas Hofmann

Presented by Yookyung Jo

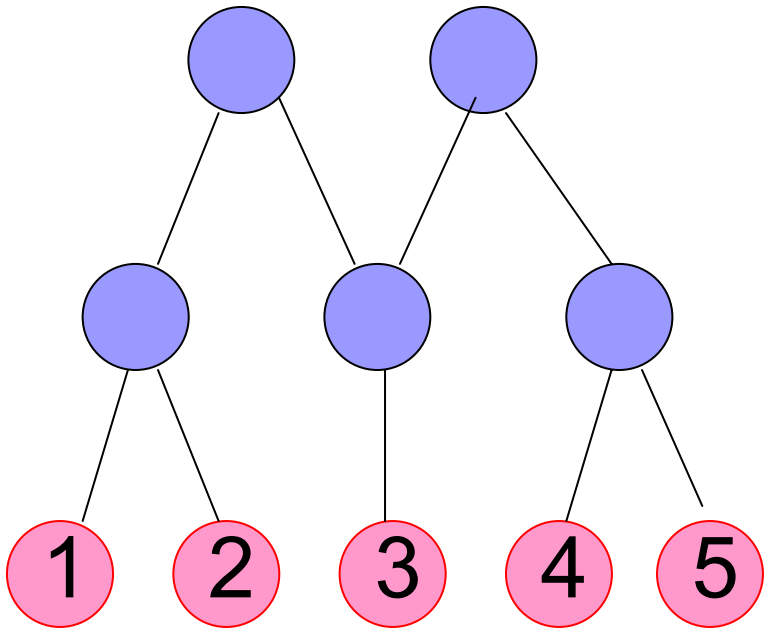
Hierarchical document classification

- Multi-class classification
  - Flat => hierarchical
- Relations among classes

  (c.f. relations among class components)

  For better learning

  For better measure of performance
- Training data for one category might be small



The ingredients

- A way to incorporate (hierarchical) proximity relation among classes
- General loss function
- SVM

Review : plain multi-class SVM

- q-class problem

$$\{(\mathrm{x}_i, y_i)\}_{i=1}^{n} \qquad 1 \le y \le q$$

$$\text{weight vector}: \quad \mathbf{W} = (\mathbf{w}_1, ..., \mathbf{w}_q)$$

$$\text{discriminant function}: \; F(\mathbf{x}, y; \mathbf{w}) \equiv \langle \mathbf{w}_y, \mathbf{x} \rangle$$

$$\text{classifier}:$$

$$f(\mathbf{x}; \mathbf{w}) \equiv \underset{y \in Y}{\operatorname{argmax}} F(\mathbf{x}, y; \mathbf{w}) = \underset{y \in Y}{\operatorname{argmax}} \langle \mathbf{w}_y, \mathbf{x} \rangle$$

$$\text{margin}: \; \gamma_i(\mathbf{w}) \equiv F(\mathbf{x}_i, y_i) - \max_{y \neq y_i} F(\mathbf{x}_i, y)$$

Review : plain multi-class SVM

- Soft-margin multi-class SVM :

$$\min_{\mathbf{w},\xi} \left(\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i \right)$$

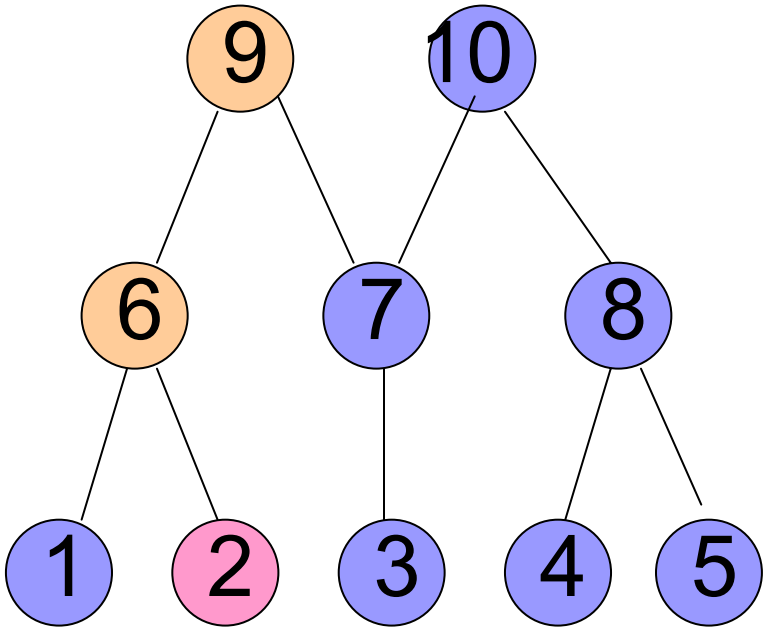
$$\text{s.t.} \quad \gamma_i(\mathbf{w}) \geq 1 - \xi_i , \quad \xi_i \geq 0 \quad (\forall i)$$

$$\text{or} \quad \left\langle \mathbf{w}_{y_i} - \mathbf{w}_y , \mathbf{x}_i \right\rangle \geq 1 - \xi_i \quad (\forall y \neq y_i) , \quad \xi_i \geq 0 \quad (\forall i)$$

Hierarchical discriminant function

flat

$$\begin{pmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \mathbf{w}_3 \\ \mathbf{w}_4 \\ \mathbf{w}_5 \end{pmatrix} \bullet \begin{pmatrix} 0 \\ \mathbf{x} \\ 0 \\ 0 \\ 0 \end{pmatrix}$$



hierarchical

$$\begin{pmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \mathbf{w}_3 \\ \mathbf{w}_4 \\ \mathbf{w}_5 \\ \mathbf{w}_6 \\ \mathbf{w}_7 \\ \mathbf{w}_8 \\ \mathbf{w}_9 \\ \mathbf{w}_{10} \end{pmatrix} \bullet \begin{pmatrix} 0 \\ \lambda_2(y)\mathbf{x} \\ 0 \\ 0 \\ 0 \\ \lambda_6(y)\mathbf{x} \\ 0 \\ 0 \\ \lambda_9(y)\mathbf{x} \\ 0 \end{pmatrix}$$

Hierarchical multi-class SVM

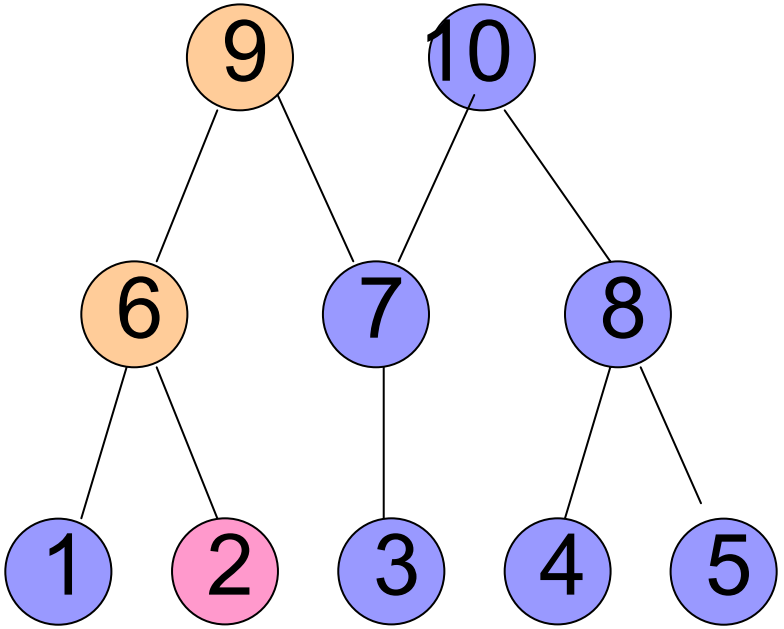
$$F(\mathbf{x}, y; \mathbf{w}) \equiv \langle \mathbf{w}, \Phi(\mathbf{x}, y) \rangle$$

$$\Phi(\mathbf{x}, y) = \begin{pmatrix} \lambda_1(y) \cdot \mathbf{x} \\ \lambda_2(y) \cdot \mathbf{x} \\ \cdots \\ \lambda_s(y) \cdot \mathbf{x} \end{pmatrix} = \Lambda(y) \otimes \mathbf{x}$$

$$\lambda_z(y) = \begin{cases} v_z, & \text{if } z \prec y \\ 0, & \text{otherwise} \end{cases}$$

$$F(\mathbf{x}, y; \mathbf{w}) = \sum_{r=1}^{s} \lambda_r(y) \langle \mathbf{w}_r, \mathbf{x} \rangle$$

Hierarchical discriminant function



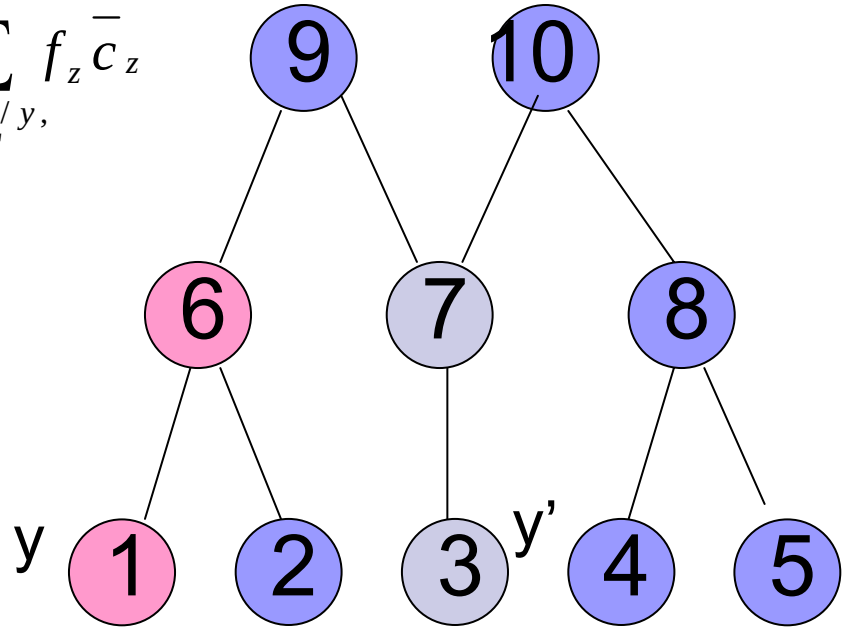
$$\Lambda(y=2)=\begin{pmatrix}0\\1\\0\\0\\0\\1\\0\\0\\1\\0\end{pmatrix},\quad\begin{pmatrix}0\\\mathbf{x}\\0\\0\\0\\\mathbf{x}\\0\\0\\\mathbf{x}\\0\end{pmatrix}=\Phi(\mathbf{x},2)$$

$$\langle \mathbf{w}, \Phi(\mathbf{x},2)\rangle = \langle \mathbf{w}_2, \mathbf{x}\rangle + \langle \mathbf{w}_6, \mathbf{x}\rangle + \langle \mathbf{w}_9, \mathbf{x}\rangle$$

Hierarchical Loss Function

- Meaningful loss function
  - Δ(y,y') : shortest path between y, y'
  - In the context of document filtering system

$$\Delta(y, y') = \sum_{\substack{z: z \prec y, \\ z \not\prec y'}} f_z c_z + \sum_{\substack{z: z \not\prec y, \\ z \prec y'}} f_z \bar{c}_z$$



Loss function in a document filtering system

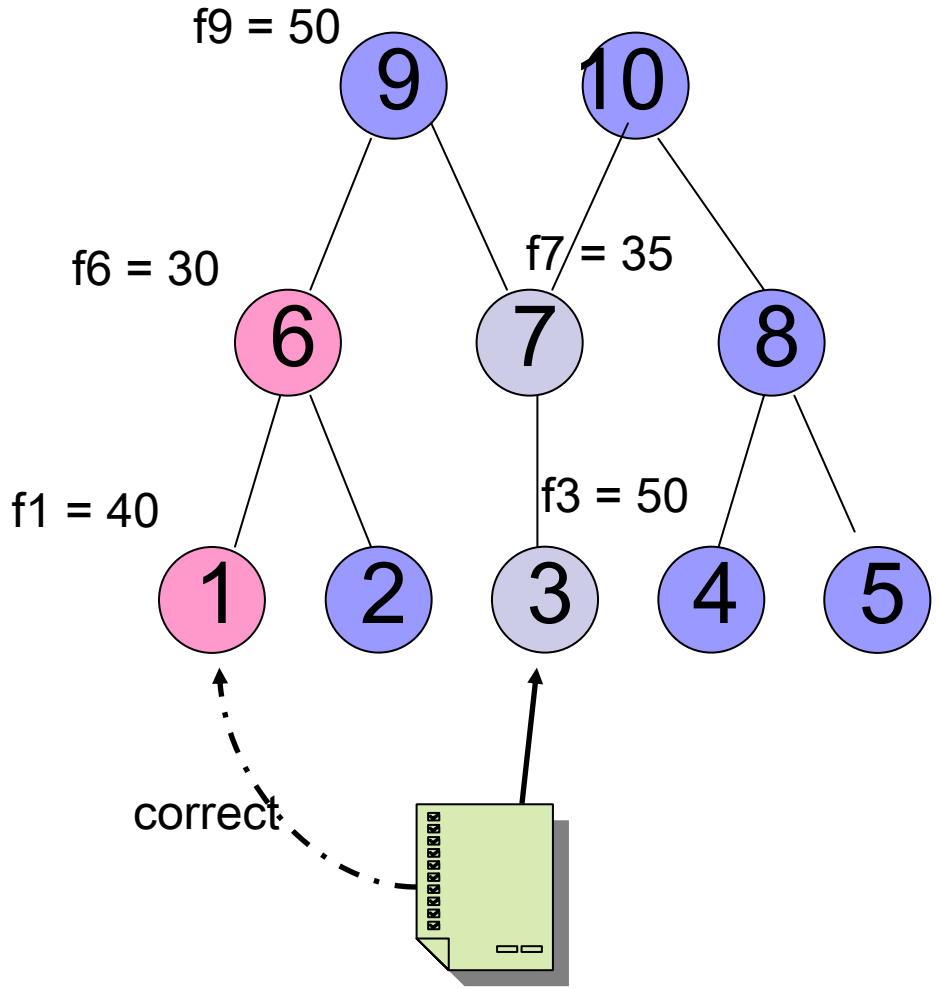
$$Loss\ (y = 1, y' = 3) =$$

$$40 * c_1 + 30 * c_6$$

miss

$$+ 50 * \bar{c}_3 + 35 * \bar{c}_7$$

overload



Hierarchical Cost-sensitive SVM

Primal Quadratic Program

$$\min_{\mathbf{w},\xi} \left(\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i \right)$$

$$\text{s.t.} \quad \left\langle \mathbf{w}, \delta\Phi_i(y) \right\rangle \geq 1 - \frac{\xi_i}{\Delta(y_i, y)}, \quad (\forall i, y \neq y_i)$$

$$\xi_i \geq 0, \quad (\forall i)$$

$$\left(\delta\Phi_i(y) \equiv \Phi(x_i, y_i) - \Phi(x_i, y)\right)$$

$$\ell(w, \xi, \alpha, \zeta) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \zeta_i \xi_i$$

$$- \sum_{i=1}^{n} \sum_{y \neq y_i} \alpha_{iy} \left(\left\langle \delta\Phi_i(y), \mathbf{w} \right\rangle - 1 + \frac{\xi_i}{\Delta(y_i, y)} \right)$$

Hierarchical Cost-sensitive SVM

Dual Quadratic program

$$\max_{\boldsymbol{\alpha}} \Theta(\boldsymbol{\alpha})$$

s.t. $\alpha_{iy} \geq 0 , \ (\forall i)$

$$\sum_{y \neq y_i} \frac{\alpha_{iy}}{\Delta(y_i, y)} \leq C , (\forall i, y \neq y_i)$$

$$\mathbf{w} = \sum_{i=1}^{n} \sum_{y \neq y_i} \alpha_{iy} \delta\Phi_i(y)$$

$$\Theta(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \sum_{y \neq y_i} \alpha_{iy} - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{y \neq y_i} \sum_{y' \neq y_j} \alpha_{iy} \alpha_{jy'} \left\langle \delta\Phi_i(y), \delta\Phi_j(y') \right\rangle$$

Efficient optimization: variable selection strategy

- Starting from a feasible solution, keep increasing the constraints from the most violated

Optimization Algorithm

Algorithm 1 Optimization algorithm using variable selection and subspace optimization.

inputs: training data $\{\mathbf{x}_i, y_i\}_{i=1}^n$, tolerance $\epsilon \geq 0$
initialize $S_i = \emptyset$, $\alpha_{iy} = 0$, for $i = 1, \ldots, n$, $y \neq y_i$
repeat
  compute F_{iy} from (28) and ψ_i from (33)
  select $\hat{i} = \text{argmax}_{i=1}^n \psi_i$
  select $\hat{y} = \text{argmax}_{y \neq y_{\hat{i}}} F_{\hat{i}y}$
  $S_{\hat{i}} = S_{\hat{i}} \cup \{\hat{y}\}$
  solve reduced QP over $\{\alpha_{\hat{i}y} : y \in S_{\hat{i}}\}$ [8a]
  solve reduced QP over $\bigcup_{i=1}^n \{\alpha_{iy} : y \in S_i\}$ [8b]
  $S_{\hat{i}} = S_{\hat{i}} - \{y : \alpha_{\hat{i},y} = 0\}$
until $\psi_{\hat{i}} \leq \epsilon$

Experimental Setup

- Synthetic Data
- WIPO-alpha collection (patent collection with taxonomy)

- Loss : hierarchical distance loss
- Kernel : linear kernel

Evaluation metric

- Accuracy :

$$\text{acc}(f) = \frac{1}{n}\sum_{i=1}^{n}\left[f(\mathbf{x}_i) = y_i\right]$$

- Precision :

$$\text{prec}(f) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\left|\{y : F(\mathbf{x}_i, y) \geq F(\mathbf{x}_i, y_i)\}\right|}\right)$$

- Taxonomy-based Loss :

$$\Delta\text{-loss}(f) = \frac{1}{n}\sum_{i=1}^{n}\Delta\left(y_i, f(\mathbf{x}_i)\right)$$

- Parent Accuracy :

$$\text{pacc}(f) = \frac{1}{n}\sum_{i=1}^{n}\left[\text{parent}\left(f(\mathbf{x}_i)\right) = \text{parent}\left(y_i\right)\right]$$

Evaluation metric : precision

$F(\mathbf{x}, y_5; \mathbf{w})$

$F(\mathbf{x}, y_3; \mathbf{w})$

$F(\mathbf{x}, y_2; \mathbf{w})$

$F(\mathbf{x}, y_1; \mathbf{w})$

$F(\mathbf{x}, y_4; \mathbf{w})$

Precision = 1/3

Experiment : Synthetic data

- Data generation :
  - Random weight vector for each node
  - Category weight vector :
    - Sum of all weight vectors along the path
  - Data points :
    - Randomly generated
    - Assigned to the category with highest score only if (highest score – second score > threshold)

Experiment : synthetic data

#children	depth	ρ	acc (%)		prec (%)		$\triangle$-loss		pacc (%)	
			flat	hsvm	flat	hsvm	flat	hsvm	flat	hsvm
3	3	0.001	68.9	72.7	81.7	84.2	0.621	0.505	80.1	84.4
		0.1	83.4	89.9	90.8	94.7	0.351	0.205	88.0	92.9
3	2	0.001	87.1	90.0	93.1	94.6	0.193	0.158	93.6	94.2
		0.1	97.4	98.7	98.7	99.3	0.0478	0.0236	97.9	99.0
6	2	0.001	67.5	69.3	80.2	82.0	0.513	0.465	81.1	84.2
		0.1	85.2	90.5	90.9	94.4	0.244	0.15	90.4	94.7

Experiment : WIPO-alpha collection

- Patent collection with taxonomy
- features from the title and claim contents

Experiment : WIPO

section	#cat	#doc	acc (%)		prec (%)		△-loss		pacc (%)	
			flat	hsvm	flat	hsvm	flat	hsvm	flat	hsvm
A	694	10962	42.3	**42.9**	51.7	**53.2**	1.24	**1.15**	61.5	**65.0**
B	1172	14690	33.2	**33.8**	41.5	**43.1**	1.54	**1.41**	57.3	**62.2**
C	852	16245	**35.5**	35.1	44.8	44.6	1.32	**1.23**	61.5	**65.6**
D	160	1710	41.8	**42.8**	52.3	**54.4**	1.20	**1.08**	65.4	**69.1**
E	230	3027	**34.7**	34.3	44.8	**46.3**	1.38	**1.30**	62.7	**64.2**
F	675	6685	31.2	**32.4**	40.6	**42.9**	1.47	**1.33**	57.6	**63.3**
G	470	10302	41.0	**41.2**	50.3	**51.1**	1.32	**1.26**	60.6	**63.0**
H	403	11629	43.0	**43.1**	54.2	**55.2**	1.12	**1.07**	63.3	**66.2**

Fewer training data

data	#cat	#doc	acc (%)		prec (%)		△-loss		pacc (%)	
			flat	hsvm	flat	hsvm	flat	hsvm	flat	hsvm
A, sample 3	694	1781	10.6	**11.7**	17.3	**20.5**	2.12	**1.87**	34.9	**43.2**
B, sample 3	1172	3033	9.56	**11.3**	14.7	**18.9**	2.25	**1.99**	36.5	**45.6**
C, sample 3	852	2212	12.1	**13.3**	18.1	**20.7**	1.90	**1.69**	45.4	**53.0**
D, sample 3	160	391	19.7	**20.5**	27.2	**30.9**	1.71	**1.54**	48.9	**57.3**
E, sample 3	230	600	10.2	**11.4**	17.3	**20.6**	2.01	**1.82**	40.5	**48.3**
F, sample 3	675	1729	13.1	**14.5**	19.4	**22.8**	2.02	**1.75**	40.8	**50.5**
G, sample 3	470	1228	12.4	**13.6**	18.9	**22.4**	2.09	**1.87**	35.2	**43.5**
H, sample 3	403	1084	14.8	**15.7**	22.6	**25.0**	1.81	**1.66**	42.0	**48.0**

Related work

- Large-margin approach
  - I.Tsochantaridis et al.,
    - Large-margin over joint feature map, slack rescaling, cutting plane algorithm
- Hierarchical document classification
  - D.Koller et al.. 1997
    - Classification => multiple classifications according to hierarchy
    - Fewer features => better learning
  - A.McCallum et al., 1998
    - Shrinkage : smoothing to overcome data sparseness
  - S.Dumais et al., 2000
    - Applying SVM to decomposed classifications

Conclusion

- SVM for hierarchical document classification
  - Framework to incorporate hierarchical relationship among classes
  - Framework with general loss function
    - an example of a meaningful loss function
  - Effective in sparse data