Hierarchical Classification via Orthogonal Transfer

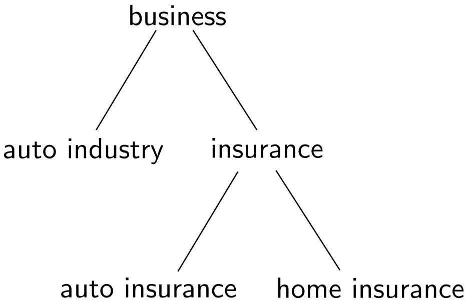
Dengyong Zhou, **Lin Xiao** and Mingrui Wu

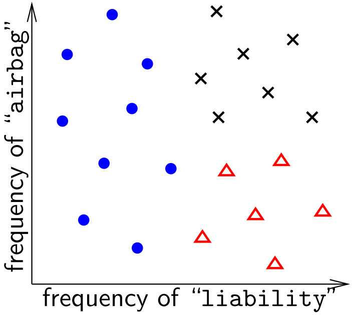
Microsoft Research

NIPS Workshop on Optimization for Machine Learning
December 10, 2010

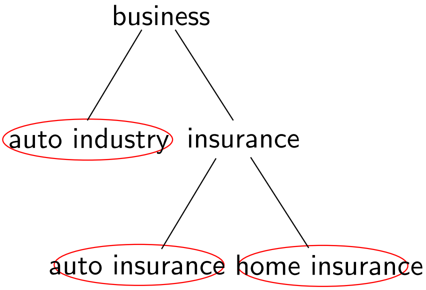
Hierarchical classification

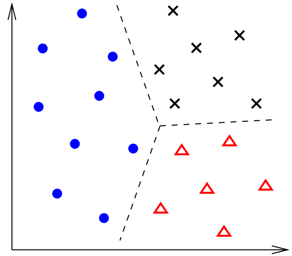
- multi-class classification with hierarchical structure
  - set of labels $\mathcal{Y} = \{1, 2, \dots, L\}$ organized as category tree
  - training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$, where $\mathbf{x}_i \in \mathbf{R}^n$, $y \in \mathcal{Y}$
  - need to learn classification function $f : \mathbf{R}^n \to \mathcal{Y}$





Flat multi-class SVM





- ignoring hierarchical structure, only consider leaf categories
- multi-class SVM (Weston & Watkins, 1999; Crammer & Singer, 2001)

$$\begin{aligned} &\text{minimize} \quad && \frac{1}{2}\sum_{v\in\mathcal{L}} \|\mathbf{w}_v\|^2 + \frac{C}{m}\sum_{i=1}^{m}\xi_i \\ &\text{subject to} \quad && \mathbf{w}_{y_i}^T\mathbf{x}_i - \mathbf{w}_v^T\mathbf{x}_i \geq 1-\xi_i,\ \forall\, v \neq y_i,\ \forall\, i=1,\ldots,m \end{aligned}$$

Exploiting hierarchical structure

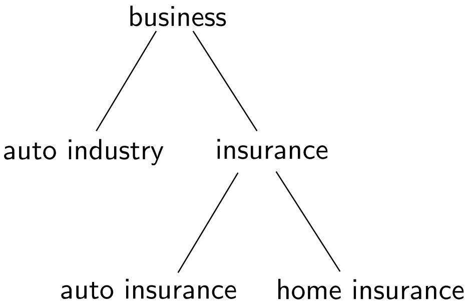
- **question:** can we improve accuracy using hierarchical structure?
- previous work
  - decomposition: solve separate, independent multi-class SVMs at each non-leaf node (Koller and Sahami, 1997; Weigend, Wiener and Pedersen, 1999; Dumais and Chen, 2000)
  - hierarchy-induced regularization: force classifiers at adjacent nodes to be similar (McCallum, Rosenfeld, Mitchell and Ng, 1998; Cai and Hofmann, 2004; Evgeniou, Micchelli and Pontil, 2005)
  - tree-induced loss: penalize classification errors between two classes by amounts proportional to their graph distance (Cai and Hofmann, 2004; Dekel, Keshet and Singer, 2004)
  - ...

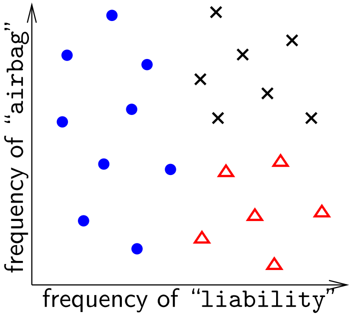
Exploiting hierarchical structure

- **question:** can we improve accuracy using hierarchical structure?
- previous work
  - decomposition: solve separate, independent multi-class SVMs at each non-leaf node (Koller and Sahami, 1997; Weigend, Wiener and Pedersen, 1999; Dumais and Chen, 2000)
  - hierarchy-induced regularization: force classifiers at adjacent nodes to be similar (McCallum, Rosenfeld, Mitchell and Ng, 1998; Cai and Hofmann, 2004; Evgeniou, Micchelli and Pontil, 2005)
  - tree-induced loss: penalize classification errors between two classes by amounts proportional to their graph distance (Cai and Hofmann, 2004; Dekel, Keshet and Singer, 2004)
  - ...

  none could really outperform flat multi-class SVM in accuracy

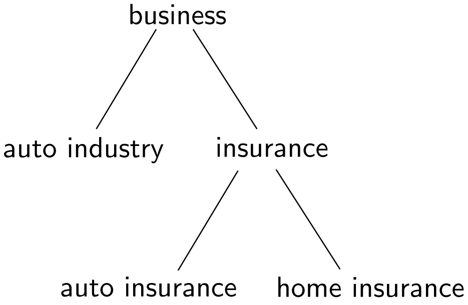
Key observation

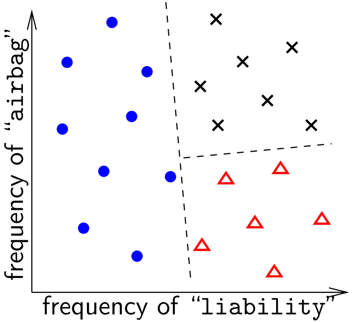




- classification at different levels of hierarchy may rely on
  - very different features
  - different combinations of same features

Key observation





- classification at different levels of hierarchy may rely on
  - very different features
  - different combinations of same features

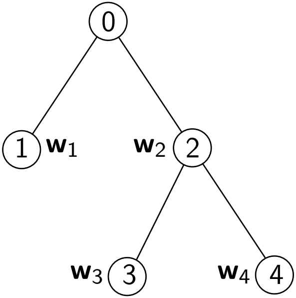
- **idea:** make classifiers at different levels different (**orthogonal**)

Outline

- hierarchical SVM with orthogonal transfer
- a sufficient condition for convexity
- optimization algorithm: regularized dual averaging method
- preliminary experiments

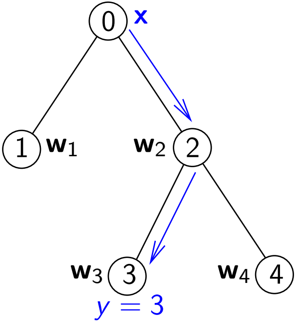
Problem setting

- some tree notations
  - $\mathcal{C}(v)$: set of children of v
  - $\mathcal{S}(v)$: set of siblings of v
  - $\mathcal{A}(v)$: set of ancestors of v
  - $\mathcal{A}^+(v) = \mathcal{A}(v) \cup \{v\}$
  - $\mathcal{D}(v)$: set of descendants of v
  - $\mathcal{D}^+(v) = \mathcal{D}(v) \cup \{v\}$



Problem setting

- some tree notations
  - $\mathcal{C}(v)$: set of children of v
  - $\mathcal{S}(v)$: set of siblings of v
  - $\mathcal{A}(v)$: set of ancestors of v
  - $\mathcal{A}^+(v) = \mathcal{A}(v) \cup \{v\}$
  - $\mathcal{D}(v)$: set of descendants of v
  - $\mathcal{D}^+(v) = \mathcal{D}(v) \cup \{v\}$



- recursive classifier

$$f(\mathbf{x}) = \left\{ \begin{array}{l} \textbf{initialize } v := 0 \\ \textbf{while } \mathcal{C}(v) \text{ is not empty} \\ \qquad v := \underset{u \in \mathcal{C}(v)}{\operatorname{argmax}} \ \mathbf{w}_u^\top \mathbf{x} \\ \textbf{return } v \end{array} \right\}$$

Hierarchical SVM with Orthogonal Transfer

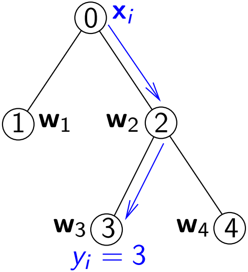
idea: make normal vectors of hyperplanes **orthogonal** to those at its ancestors as much as possible

$$\text{minimize} \quad \frac{1}{2}\sum_{v\in\mathcal{Y}} k(v,v)\|\mathbf{w}_v\|^2 + \sum_{v\in\mathcal{Y}}\sum_{u\in\mathcal{A}(v)} k(u,v)|\mathbf{w}_u^\top \mathbf{w}_v| + \frac{C}{m}\sum_{i=1}^{m}\xi_i$$

$$\text{subject to} \quad \mathbf{w}_v^\top \mathbf{x}_i - \mathbf{w}_u^\top \mathbf{x}_i \geq 1 - \xi_i,\ \forall\, u \in \mathcal{S}(v),\ \forall\, v \in \mathcal{A}^+(y_i),$$

$$\xi_i \geq 0, \quad i = 1, \ldots, m$$

- penalizing $|\mathbf{w}_u^T \mathbf{w}_v|$ encourages orthogonality
- $k(u,v) \geq 0$ are given parameters
- $(\mathbf{x}_i, y_i)$ used for discriminating y_i and its ancestors from their own siblings
- classifiers that are not siblings never appear in same constraint



A sufficient condition for convexity

- it suffices to establish convexity of

$$\Omega(\mathbf{w}) = \sum_{v \in \mathcal{Y}} k(v, v) \|\mathbf{w}_v\|^2 + \sum_{u \neq v} k(u, v) |\mathbf{w}_u^\top \mathbf{w}_v|$$

- **Theorem.** If the symmetric matrix $\bar{k}$ defined by

$$\bar{k}(u, v) = \begin{cases} k(v, v) & \text{if} \quad u = v, \\ -|k(u, v)| & \text{otherwise} \end{cases}$$

  is positive semidefinite, then $\Omega(\mathbf{w})$ is convex

- for example, $\Omega(\mathbf{w})$ convex if $k(u, v)$ is diagonally dominant

- $\Omega(\mathbf{w})$ strictly convex if $\bar{k}$ positive definite

A sufficient condition for convexity

proof idea: directly use definition of convexity

- for any two vectors $\mathbf{s}, \mathbf{t} \in \mathbf{R}^{nL}$ and any $\alpha \in [0, 1]$,

$$
\begin{aligned}
& \alpha\Omega(\mathbf{s}) + (1-\alpha)\Omega(\mathbf{t}) - \Omega\big(\alpha\mathbf{s} + (1-\alpha)\mathbf{t}\big) \\
\geq \;& \cdots \\
\geq \;& \alpha(1-\alpha)\Big(\textstyle\sum_u k(u,u)\|\mathbf{s}_u - \mathbf{t}_u\|^2 - \sum_{u\neq v} k(u,v)\|\mathbf{s}_u - \mathbf{t}_u\|\|\mathbf{s}_v - \mathbf{t}_v\|\Big) \\
= \;& \alpha(1-\alpha)\textstyle\sum_{u,v} \bar{k}(u,v)\|\mathbf{s}_u - \mathbf{t}_u\|\|\mathbf{s}_v - \mathbf{t}_v\| \\
\geq \;& 0
\end{aligned}
$$

- in fact, we only need $\bar{k}$ to be *copositive*

Representer theorem

- **Theorem.** If $\bar{k}$ is positive definite, then the solution admits a representation of the form

$$\mathbf{w}_v = \sum_{i=1}^{m} c_{vi} \mathbf{x}_i, \qquad \forall\, v \in \mathcal{Y}$$

- possible to extend in more general reproducing kernel Hilbert space (RKHS) with nonlinear classifiers

How to solve it efficiently?

- hierarchical SVM with orthogonal transfer

$$
\begin{aligned}
\text{minimize} \quad & \frac{1}{2}\sum_{v\in\mathcal{Y}} k(v,v)\|\mathbf{w}_v\|^2 + \sum_{v\in\mathcal{Y}}\sum_{u\in\mathcal{A}(v)} k(u,v)|\mathbf{w}_u^\top \mathbf{w}_v| + \frac{C}{m}\sum_{i=1}^{m}\xi_i \\
\text{subject to} \quad & \mathbf{w}_v^\top \mathbf{x}_i - \mathbf{w}_u^\top \mathbf{x}_i \geq 1 - \xi_i,\ \forall\, u \in \mathcal{S}(v),\ \forall\, v \in \mathcal{A}^+(y_i), \\
& \xi_i \geq 0, \quad i = 1, \ldots, m.
\end{aligned}
$$

How to solve it efficiently?

- hierarchical SVM with orthogonal transfer

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\sum_{v\in\mathcal{Y}} k(v,v)\|\mathbf{w}_v\|^2 + \sum_{v\in\mathcal{Y}}\sum_{u\in\mathcal{A}(v)} k(u,v)|\mathbf{w}_u^\top \mathbf{w}_v| + \frac{C}{m}\sum_{i=1}^{m}\xi_i \\ \text{subject to} \quad & \mathbf{w}_v^\top \mathbf{x}_i - \mathbf{w}_u^\top \mathbf{x}_i \geq 1-\xi_i,\ \forall\, u\in\mathcal{S}(v),\ \forall\, v\in\mathcal{A}^+(y_i), \\ & \xi_i \geq 0, \quad i=1,\ldots,m. \end{aligned}$$

- an equivalent unconstrained optimization problem

$$\begin{aligned} \underset{\mathbf{w}\in\mathbf{R}^{nL}}{\text{minimize}} \quad & \frac{1}{2}\sum_{v\in V} k(v,v)\|\mathbf{w}_v\|^2 + \sum_{v\in V}\sum_{u\in A(v)} k(u,v)\left|\mathbf{w}_u^T \mathbf{w}_v\right| \\ & + \frac{C}{m}\sum_{i=1}^{m} \max\left\{0, \max_{u\in S(v), v\in A^+(y_i)} \left\{1-\mathbf{w}_v^T\mathbf{x}_i + \mathbf{w}_u^T\mathbf{x}_i\right\}\right\} \end{aligned}$$

Splitting objective function

- define

$$J(\mathbf{w}) \triangleq \frac{1}{2}\sum_{v\in V} k(v,v)\|\mathbf{w}_v\|^2 + \sum_{v\in V}\sum_{u\in A(v)} k(u,v)\left|\mathbf{w}_u^T\mathbf{w}_v\right| + \cdots$$

- assume $\bar{k}$ positive definite, and $\lambda_{\min}(\bar{k})$ is smallest eigenvalue
- split as $J(\mathbf{w}) = \phi(\mathbf{w}) + \Psi(\mathbf{w})$

$$\begin{aligned}\phi(\mathbf{w}) &= \frac{1}{2}\sum_{v\in\mathcal{Y}}\big(k(v,v) - \lambda_{\min}\big)\|\mathbf{w}_v\|^2 + \sum_{v\in V}\sum_{u\in A(v)} k(u,v)\left|\mathbf{w}_u^T\mathbf{w}_v\right| \\ &\quad + \cdots \\ \Psi(\mathbf{w}) &= \frac{\lambda_{\min}}{2}\sum_{v\in\mathcal{Y}}\|\mathbf{w}_v\|^2 = \frac{\lambda_{\min}}{2}\|\mathbf{w}\|^2\end{aligned}$$

Optimization of composite objective

consider generic optimization problem

$$\underset{\mathbf{w}\in\mathcal{W}}{\text{minimize}} \quad J(\mathbf{w}) = \phi(\mathbf{w}) + \Psi(\mathbf{w})$$

- ϕ convex, possibly nonsmooth
- Ψ strongly convex with convexity parameter σ
- Ψ simple: given $\mathbf{g}$, it is easy to solve

$$\underset{\mathbf{w}\in\mathcal{W}}{\text{minimize}} \quad \langle \mathbf{g}, \mathbf{w} \rangle + \Psi(\mathbf{w})$$

if $\mathbf{g}$ is subgradient of ϕ, this also gives a lower bound of $J(\mathbf{w}^\star)$

Regularized dual averaging method

input: $\epsilon > 0$
initialize: $t = 1$, $\mathbf{w}(1) = 0$, $\bar{\mathbf{g}}(0) = 0$, $\overline{J}(1) = C$, $\underline{J}(1) = 0$
repeat

1. *compute* $\mathbf{g}(t) \in \partial\phi(\mathbf{w}(t))$, *and* $\bar{\mathbf{g}}(t) = \dfrac{t-1}{t}\bar{\mathbf{g}}(t-1) + \dfrac{1}{t}\mathbf{g}(t)$
2. *compute* $\mathbf{w}(t+1) = \underset{\mathbf{w}\in\mathcal{W}}{\operatorname{argmin}} \left\{\bar{\mathbf{g}}(t)^\top \mathbf{w} + \Psi(\mathbf{w})\right\}$
3. *update upper bound* $\overline{J}(t+1)$ *and lower bound* $\underline{J}(t+1)$

until $\overline{J}(t+1) - \underline{J}(t+1) \leq \epsilon$

Regularized dual averaging method

input: $\epsilon > 0$
initialize: $t = 1$, $\mathbf{w}(1) = 0$, $\bar{\mathbf{g}}(0) = 0$, $\overline{J}(1) = C$, $\underline{J}(1) = 0$
repeat

1. *compute* $\mathbf{g}(t) \in \partial\phi(\mathbf{w}(t))$, *and* $\bar{\mathbf{g}}(t) = \dfrac{t-1}{t}\bar{\mathbf{g}}(t-1) + \dfrac{1}{t}\mathbf{g}(t)$
2. *compute* $\mathbf{w}(t+1) = \operatorname*{argmin}_{\mathbf{w}\in\mathcal{W}} \left\{\bar{\mathbf{g}}(t)^\top \mathbf{w} + \Psi(\mathbf{w})\right\}$
3. *update upper bound* $\overline{J}(t+1)$ *and lower bound* $\underline{J}(t+1)$

until $\overline{J}(t+1) - \underline{J}(t+1) \leq \epsilon$

- extends Nesterov (2005) to strongly convex functions
- iteration complexity: $O\left(\ln(t)/t\right)$
  (c.f. Pegasos by Shalev-Shwartz, Singer and Srebro, 2007)

Preliminary experiments

- text categorization data sets:
  - RCV1-v2 (Lewis, Yang, Rose and Li, 2004)
  - 20-newsgroups
- all formulations solved by regularized dual average method
- classification error rates on testing sets (average performance over 50 rounds of random splitting of training/testing datasets)

Methods	MACT	CCAT	ECAT	20-news
Flat Multiclass SVM	5.26($\pm$0.20)	21.49($\pm$0.27)	11.85($\pm$0.29)	11.50($\pm$0.50)
Hier. Multiclass SVM	4.87($\pm$0.18)	21.48($\pm$0.31)	12.09($\pm$0.34)	11.37($\pm$0.49)
Hier. Multitask SVM	4.73($\pm$0.18)	21.99($\pm$0.32)	12.05($\pm$0.33)	11.36($\pm$0.48)
Hier. SVM (path loss)	13.55($\pm$0.60)	26.48($\pm$0.42)	15.40($\pm$0.43)	33.22($\pm$1.14)
Hier. SVM (0/1 loss)	6.65($\pm$0.22)	22.21($\pm$0.31)	13.01($\pm$0.32)	11.95($\pm$0.54)
Orthogonal Transfer	**3.03**($\pm$**0.13**)	**17.53**($\pm$**0.55**)	**10.01**($\pm$**0.28**)	**11.19**($\pm$**0.46**)

Summary

- hierarchical classification via orthogonal transfer
  - key observation: classification at different levels often rely on different features, or different combinations of same features
  - convexity condition for orthogonality regularization
  - preliminary experiments very promising

Summary

- hierarchical classification via orthogonal transfer
  - key observation: classification at different levels often rely on different features, or different combinations of same features
  - convexity condition for orthogonality regularization
  - preliminary experiments very promising

- results of independent interests:
  - optimizing for orthogonality

$$\underset{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{W}}{\text{minimize}} \quad \sum_{i,j} k_{ij} \left|\mathbf{w}_i^T \mathbf{w}_j\right|$$

  - new variant of regularized dual averaging method

Summary

- hierarchical classification via orthogonal transfer
  - key observation: classification at different levels often rely on different features, or different combinations of same features
  - convexity condition for orthogonality regularization
  - preliminary experiments very promising

- results of independent interests:
  - optimizing for orthogonality

$$\underset{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{W}}{\text{minimize}} \quad \sum_{i,j} k_{ij}\,|\mathbf{w}_i^T\mathbf{w}_j|$$

  - new variant of regularized dual averaging method

future work

- extensive computational experiments
- analysis from learning theory perspective