Regularized Multi-Class Semi-Supervised Boosting

Amir Saffari, Christian Leistner, Horst Bischof

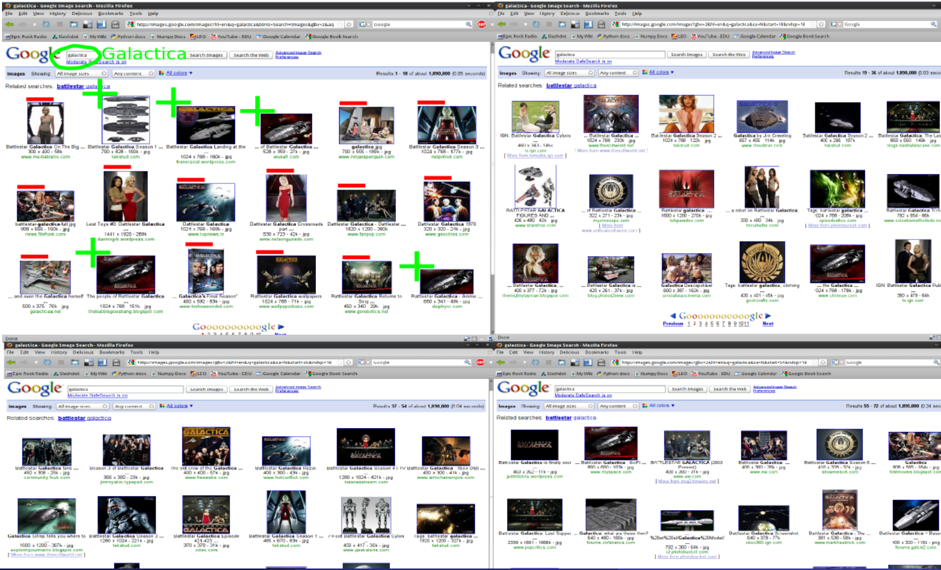
Institute for Computer Graphics and Vision, Graz University of Technology, Austria

CVPR 2009, June 22, 2009

Supervised Learning

Graz University of Technology

Semi-Supervised Learning (SSL)



Large-Scale Applications and Semi-Supervised Learning

Picasa

flickr

ImageShack

YouTube
Broadcast Yourself

facebook

myspace
a place for friends

TU Graz
Graz University of Technology

Conclusions: Beta version 0.1

- We propose a semi-supervised boosting algorithm

Graz University of Technology

Conclusions: Beta version 0.1

- We propose a semi-supervised boosting algorithm
- which solves multi-class problems without decomposing them into binary tasks.

Graz University of Technology

Conclusions: Beta version 0.1

- We propose a semi-supervised boosting algorithm
- which solves multi-class problems without decomposing them into binary tasks.
- Additionally, our algorithm scales very well with respect to the number of both labeled and unlabeled samples.

Graz University of Technology

Outline

SSL Methods

Semi-Supervised Learning

Semi-supervised learning is a class of machine learning techniques that make use of both labeled and unlabeled data for training.

- There exists many SSL methods, see:
  - X. Zhu, "Semi-Supervised Learning Survey", 2008 and
  - O. Chapelle, B. Schoelkopf, A. Zien, "The Semi-Supervised Learning", Cambridge, 2006.

Graz University of Technology

Motivations

- Many successful SSL methods do not scale very well w.r.t. the number of unlabeled samples, or are very sensitive to the choice of hyper-parameters (G. Mann, A. McCallum, ICML 2007). Expect to see $\mathcal{O}(n^3)$ many times.

Graz University of Technology

Motivations

- Many successful SSL methods do not scale very well w.r.t. the number of unlabeled samples, or are very sensitive to the choice of hyper-parameters (G. Mann, A. McCallum, ICML 2007). Expect to see $\mathcal{O}(n^3)$ many times.
- Usually multi-class problems are solved via 1-vs-all and occasionally with 1-vs-1 decompositions.

Graz University of Technology

What is wrong with 1-vs-all?

- Do you want to repeat a slow method a few more of times?

Graz University of Technology

What is wrong with 1-vs-all?

- Do you want to repeat a slow method a few more of times?
- Calibration problems (B. Schoelkopf, A. Smola, 2002).

Graz University of Technology

What is wrong with 1-vs-all?

- Do you want to repeat a slow method a few more of times?
- Calibration problems (B. Schoelkopf, A. Smola, 2002).
- Artificial unbalanced binary problems.

What is wrong with 1-vs-all?

- Do you want to repeat a slow method a few more of times?
- Calibration problems (B. Schoelkopf, A. Smola, 2002).
- Artificial unbalanced binary problems.

- There exists slow multi-class SSL methods, see the details in the paper.

Graz University of Technology

Multi-Class Semi-Supervised Boosting

Multi-class classifier: $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), \cdots, f_K(\mathbf{x})]^T$.

Graz University of Technology

Multi-Class Semi-Supervised Boosting

Multi-class classifier: $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), \cdots, f_K(\mathbf{x})]^T$.

Overall Loss

$$\mathcal{L}(\mathbf{f}(\mathbf{x}), \mathcal{X}) = \underbrace{\sum_{(\mathbf{x},y)\in\mathcal{X}_l} \ell(\mathbf{f}(\mathbf{x}))}_{\text{Labeled}} + \underbrace{\alpha \sum_{\mathbf{x}\in\mathcal{X}_u} \ell_c(\mathbf{f}(\mathbf{x})) + \beta \sum_{\mathbf{x}\in\mathcal{X}_u} \ell_m(\mathbf{f}(\mathbf{x}))}_{\text{Unlabeled}} \tag{1}$$

Multi-Class Semi-Supervised Boosting

Multi-class classifier: $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), \cdots, f_K(\mathbf{x})]^T$.

Overall Loss

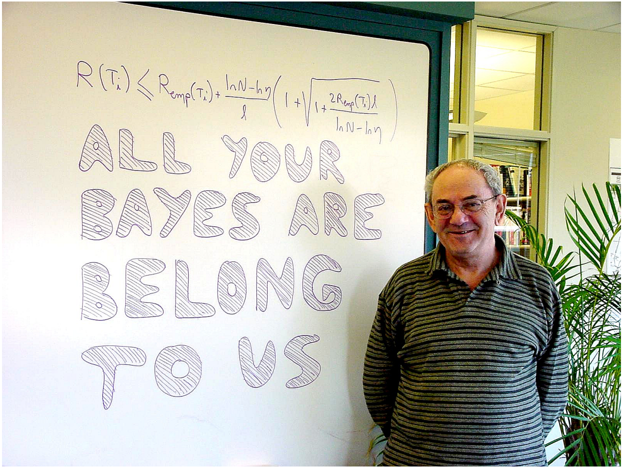
$$\mathcal{L}(\mathbf{f}(\mathbf{x}), \mathcal{X}) = \underbrace{\sum_{(\mathbf{x},y)\in\mathcal{X}_l} \ell(\mathbf{f}(\mathbf{x}))}_{\text{Labeled}} + \underbrace{\alpha \sum_{\mathbf{x}\in\mathcal{X}_u} \ell_c(\mathbf{f}(\mathbf{x})) + \beta \sum_{\mathbf{x}\in\mathcal{X}_u} \ell_m(\mathbf{f}(\mathbf{x}))}_{\text{Unlabeled}} \tag{1}$$

Boosting Model

$$\mathbf{f}(\mathbf{x}) = \nu \sum_{t=1}^{T} \mathbf{g}^t(\mathbf{x}) \tag{2}$$

Graz University of Technology

Fisher-Consistent Loss Functions



Vladimir Vapnik (picture courtesy of Yann LeCun)

Fisher-Consistent Loss Functions

Margin Vector

$\mathbf{f}(\mathbf{x})$ is a universal margin vector, if $\forall \mathbf{x} : \sum_{i=1}^{K} f_i(\mathbf{x}) = 0$.

Graz University of Technology

Fisher-Consistent Loss Functions

Margin Vector

$\mathbf{f}(\mathbf{x})$ is a universal margin vector, if $\forall \mathbf{x} : \sum_{i=1}^{K} f_i(\mathbf{x}) = 0$.

Fisher-Consistent Loss

$\ell(\cdot)$ is Fisher-consistent, if the minimization of the expected risk:

$$\hat{\mathbf{f}}(\mathbf{x}) = \arg\min_{\mathbf{f}(\mathbf{x})} \int_{(\mathbf{x},y)} \ell(f_y(\mathbf{x}))p(y,\mathbf{x})d(\mathbf{x},y) \tag{3}$$

has a unique solution and

$$C(\mathbf{x}) = \arg\max_i \hat{f}_i(\mathbf{x}) = \arg\max_i p(y = i|\mathbf{x}). \tag{4}$$

Graz University of Technology

Fisher-Consistent Loss Functions

Margin Vector

$\mathbf{f}(\mathbf{x})$ is a universal margin vector, if $\forall \mathbf{x} : \sum_{i=1}^{K} f_i(\mathbf{x}) = 0$.

Fisher-Consistent Loss

$\ell(\cdot)$ is Fisher-consistent, if the minimization of the expected risk:

$$\hat{\mathbf{f}}(\mathbf{x}) = \arg\min_{\mathbf{f}(\mathbf{x})} \int_{(\mathbf{x},y)} \ell(f_y(\mathbf{x}))p(y,\mathbf{x})d(\mathbf{x},y) \tag{3}$$

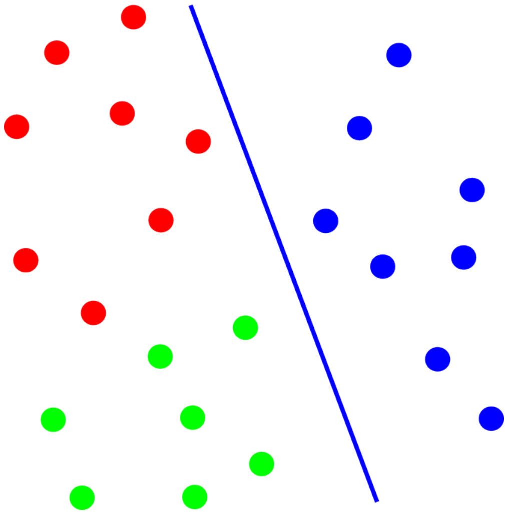
has a unique solution and

$$C(\mathbf{x}) = \arg\max_i \hat{f}_i(\mathbf{x}) = \arg\max_i p(y = i|\mathbf{x}). \tag{4}$$

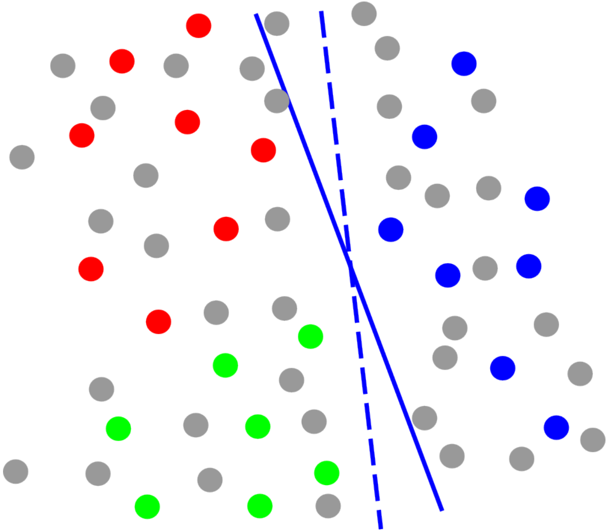
$$\mathcal{L}(\mathbf{f}(\mathbf{x}), \mathcal{X}_l) = \sum_{(\mathbf{x},y)\in\mathcal{X}_l} e^{-f_y(\mathbf{x})}$$

Graz University of Technology

Margin Assumption



Margin Assumption



Margin Assumption

Put the decision boundary over low-density regions of features space. This is equivalent to maximizing the margin of the unlabeled samples.

Example

Transductive Support Vector Machines (TSVM, T. Joachims, ICML 1999) uses this loss function for the binary SVM classifier $h(\mathbf{x})$

$$\ell_u(h(\mathbf{x})) = \max(0, 1 - |h(\mathbf{x})|) \quad (5)$$

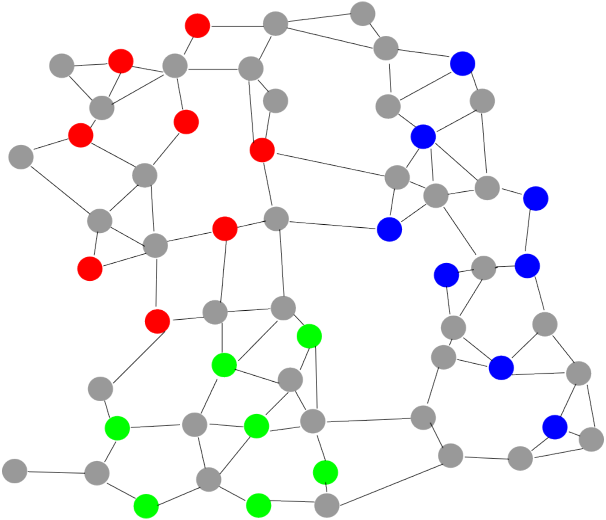
Multi-Class Unlabeled Margin

We propose to maximize the multi-class margin of the unlabeled samples by using

$$\ell_m(\mathbf{f}(\mathbf{x})) = \max(0, M - \max_i (f_i(\mathbf{x}))). \quad (6)$$

Graz University of Technology

Manifold Assumption



Graz University of Technology

Manifold Assumption

Enforce the classifier to predict similar labels for similar unlabeled samples.

Example

Graph-based methods, such as Laplacian SVM (Belkin et al., JMLR 2006), use this loss function for the binary SVM classifier $h(\mathbf{x})$

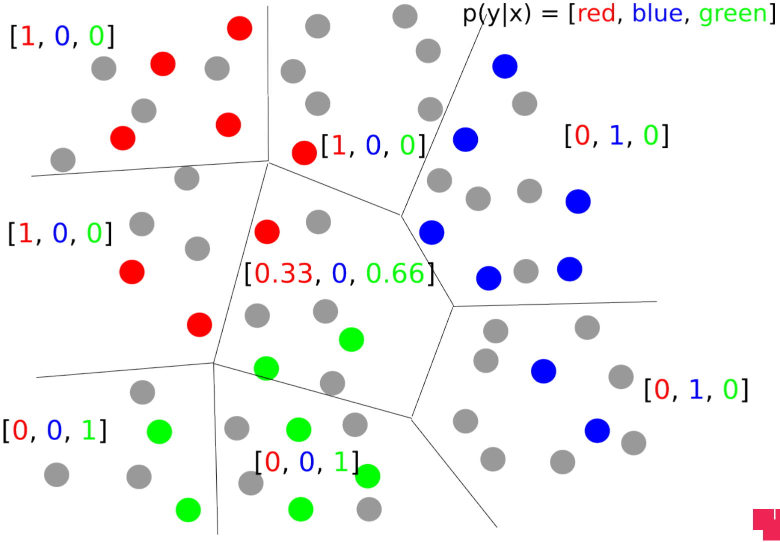
$$\ell_u(h(\mathbf{x})) = \sum_{\mathbf{x}' \in \mathcal{X}_u, \mathbf{x}' \neq \mathbf{x}} s(\mathbf{x}, \mathbf{x}') \| h(\mathbf{x}) - h(\mathbf{x}') \|^2. \tag{7}$$

Cluster Prior

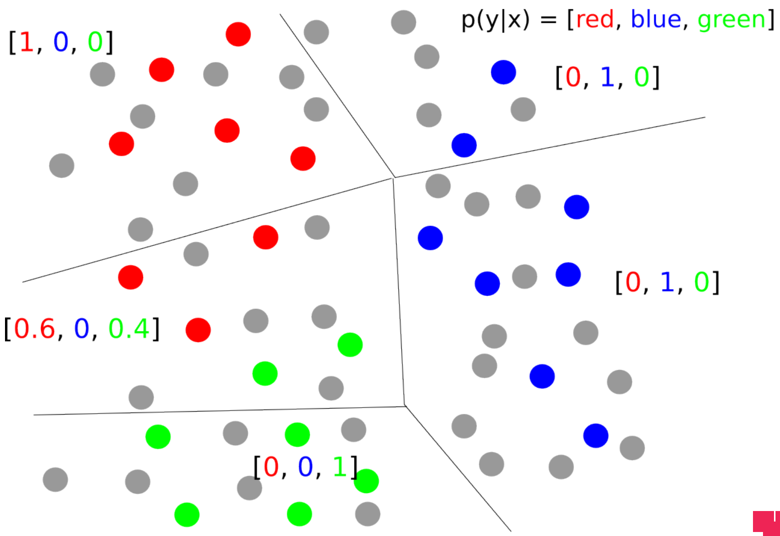
We enforce the multi-class classifier to have a consistent probabilistic estimates over regions of feature space formed by similar samples, i.e. clusters.

Graz University of Technology

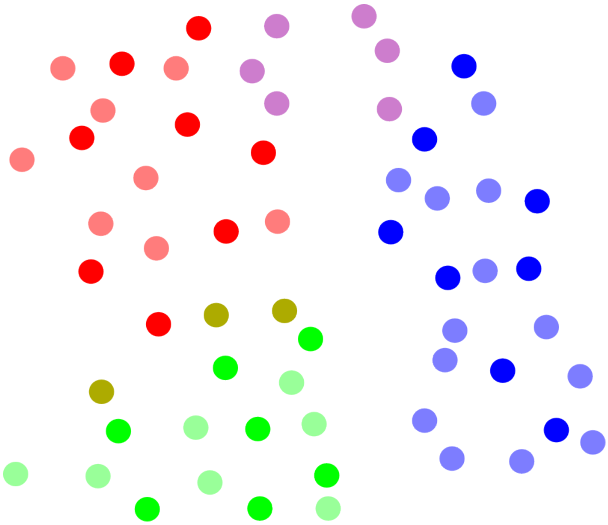
Cluster Priors



Cluster Priors



Cluster Priors



Cluster Priors

Cluster Prior

$\forall \mathbf{x} \in \mathcal{X}_u, \forall i \in \{1, \cdots, K\} : p_p(y = i|\mathbf{x})$.

We use the Kullback-Leibler (KL) divergence

$$\ell_c(\mathbf{f}(\mathbf{x})) = -\mathbf{p}_p^T \mathbf{f}(\mathbf{x}) + \log \sum_{j=1}^{K} e^{f_j(\mathbf{x})}. \tag{8}$$

Graz University of Technology

Cluster Priors

Cluster Prior

$\forall \mathbf{x} \in \mathcal{X}_u, \forall i \in \{1, \cdots, K\} : p_p(y = i|\mathbf{x})$.

We use the Kullback-Leibler (KL) divergence

$$\ell_c(\mathbf{f}(\mathbf{x})) = -\mathbf{p}_p^T \mathbf{f}(\mathbf{x}) + \log \sum_{j=1}^{K} e^{f_j(\mathbf{x})}. \tag{8}$$

- Use any clustering method which suits your application.
- Use similarity functions if it helps clustering to recover the manifolds.

Graz University of Technology

Cluster Priors

Cluster Prior

$\forall \mathbf{x} \in \mathcal{X}_u, \forall i \in \{1, \cdots, K\} : p_p(y = i|\mathbf{x})$.

We use the Kullback-Leibler (KL) divergence

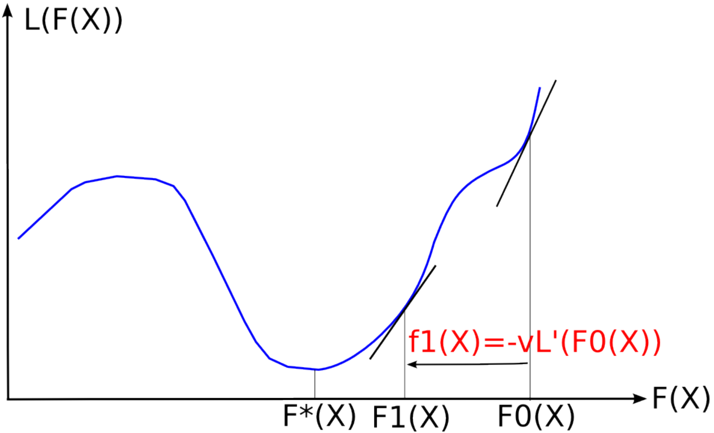
$$\ell_c(\mathbf{f}(\mathbf{x})) = -\mathbf{p}_p^T \mathbf{f}(\mathbf{x}) + \log \sum_{j=1}^{K} e^{f_j(\mathbf{x})}. \tag{8}$$

- Use any clustering method which suits your application.
- Use similarity functions if it helps clustering to recover the manifolds.
- Use any other source of information in form of priors: label prior, knowledge transfer, human prior knowledge.

Graz University of Technology

Learning with Functional Gradient Descent

$$F^*(\mathbf{x}) = F_0(\mathbf{x}) - \nu \sum_{t=1}^{T} \frac{\partial L}{\partial F}\Big|_{(F_{t-1}(\mathbf{x}))}$$



Friedman et al., Annals of Applied Statistics, 2001

RMSBoost

Learning task for t^{th} boosting stage becomes

$$\mathbf{g}^t(\mathbf{x}) = \arg\max_{\mathbf{g}(\mathbf{x})} \sum_{(\mathbf{x},y)\in\mathcal{X}_l} e^{-f_y(\mathbf{x})}\mathbf{y}^T\mathbf{g}(\mathbf{x}) + \sum_{\mathbf{x}\in\mathcal{X}_u} \left(\alpha\Delta\mathbf{p} + \beta\mathbf{m}\right)^T\mathbf{g}(\mathbf{x}). \quad (9)$$

Graz University of Technology

RMSBoost

Learning task for t^{th} boosting stage becomes

$$\mathbf{g}^t(\mathbf{x}) = \arg\max_{\mathbf{g}(\mathbf{x})} \sum_{(\mathbf{x},y)\in\mathcal{X}_l} e^{-f_y(\mathbf{x})} \mathbf{y}^T \mathbf{g}(\mathbf{x}) + \sum_{\mathbf{x}\in\mathcal{X}_u} (\alpha\Delta\mathbf{p} + \beta\mathbf{m})^T \mathbf{g}(\mathbf{x}). \quad (9)$$

Theorem

The solution using a multi-class classifier $C(\mathbf{x}) \in \{1, \cdots, K\}$ *is*

$$C_t(\mathbf{x}) = \arg\min_{C(\mathbf{x})} \sum_{(\mathbf{x},y)\in\mathcal{X}_l} w_l \mathbb{I}(C(\mathbf{x}) \neq y) + \sum_{\mathbf{x}\in\mathcal{X}_u} w_u \mathbb{I}(C(\mathbf{x}) \neq z) \quad (10)$$

where $w_l = e^{-f_y(\mathbf{x})}$ *is the weight for a labeled sample,* $z = \arg\max_i (\alpha\Delta p_i + \beta m_i)$ *and* $w_u = \alpha\Delta p_z + \beta m_z$ *are the pseudo-label and weight for an unlabeled sample, respectively.*

Experimental Settings

- RMSBoost is compared with:
  - AdaBoost.ML (Zou et al., Annals of Applied Statistics 2008)
  - Kernel SVM
  - Multi-Switch TSVM (Sindhwani and Keerthi, SIGIR 2006)
  - SERBoost (Saffari et al., ECCV 2008)
  - RMBoost

Experimental Settings

- RMSBoost is compared with:
  - AdaBoost.ML (Zou et al., Annals of Applied Statistics 2008)
  - Kernel SVM
  - Multi-Switch TSVM (Sindhwani and Keerthi, SIGIR 2006)
  - SERBoost (Saffari et al., ECCV 2008)
  - RMBoost
- Base learners are tiny extremely randomized forests, each consisting of 10 trees.

Graz University of Technology

Experimental Settings

- RMSBoost is compared with:
  - AdaBoost.ML (Zou et al., Annals of Applied Statistics 2008)
  - Kernel SVM
  - Multi-Switch TSVM (Sindhwani and Keerthi, SIGIR 2006)
  - SERBoost (Saffari et al., ECCV 2008)
  - RMBoost
- Base learners are tiny extremely randomized forests, each consisting of 10 trees.
- Boosting iterations set to be 10000.

Experimental Settings

- RMSBoost is compared with:
  - AdaBoost.ML (Zou et al., Annals of Applied Statistics 2008)
  - Kernel SVM
  - Multi-Switch TSVM (Sindhwani and Keerthi, SIGIR 2006)
  - SERBoost (Saffari et al., ECCV 2008)
  - RMBoost
- Base learners are tiny extremely randomized forests, each consisting of 10 trees.
- Boosting iterations set to be 10000.
- Results of hierarchical k-means is averaged 10 times to estimate the cluster priors.

Graz University of Technology

Experimental Settings

- RMSBoost is compared with:
  - AdaBoost.ML (Zou et al., Annals of Applied Statistics 2008)
  - Kernel SVM
  - Multi-Switch TSVM (Sindhwani and Keerthi, SIGIR 2006)
  - SERBoost (Saffari et al., ECCV 2008)
  - RMBoost
- Base learners are tiny extremely randomized forests, each consisting of 10 trees.
- Boosting iterations set to be 10000.
- Results of hierarchical k-means is averaged 10 times to estimate the cluster priors.
- All boosting and RF methods are implemented in C++ and use ATLAS subroutines.

Graz University of Technology

Machine Learning Datasets

5% of the training data is chosen randomly to form the labeled set, the rest 95% is used as unlabeled set.

Dataset	# Train	# Test	# Class	# Feat.
Letter	15000	5000	26	16
SensIt (com)	78823	19705	3	100

Table: Data sets for the machine learning experiments.

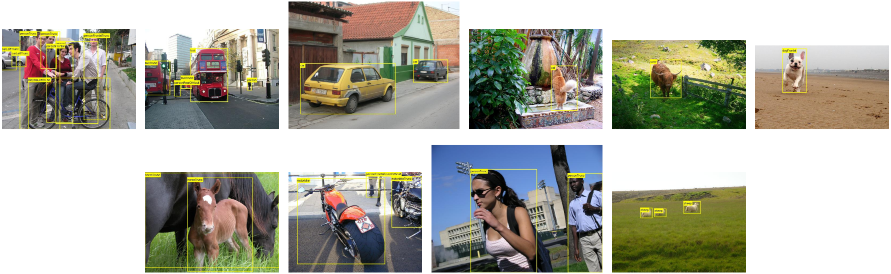
Method	AML	SVM	TSVM	SER	RMB	RMSB
Letter	72.3	70.3	65.9	76.5	74.4	**79.9**
SensIt	79.5	80.2	79.9	81.9	79.0	**83.7**

Table: Classification accuracy (in %).

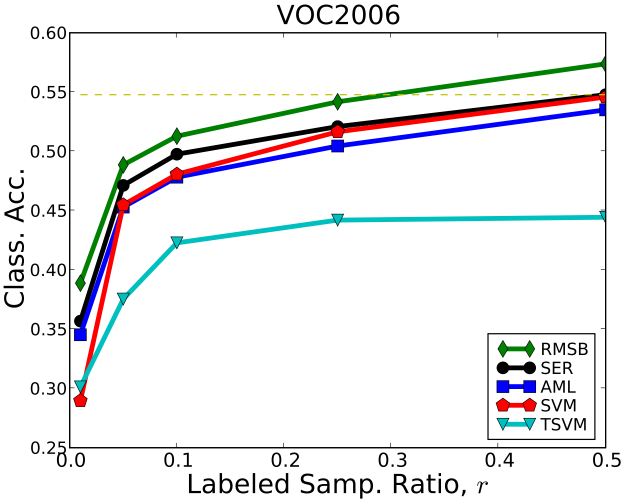
Graz University of Technology

PASCAL 2006 Object Categorization Dataset

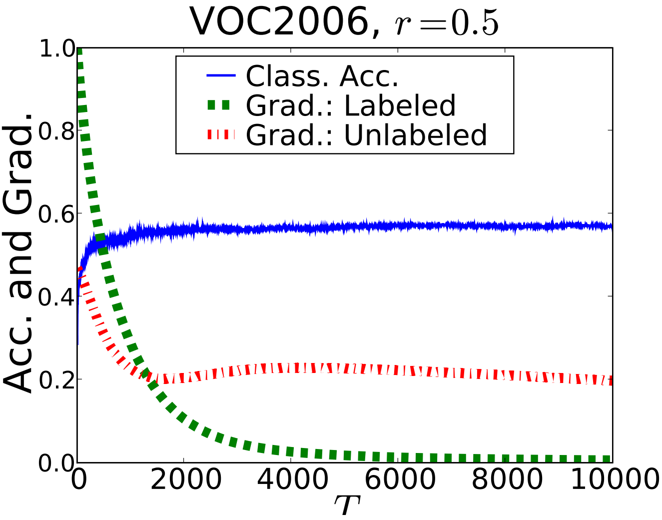
- Standard bag-of-words using quantized SIFT on a regular grid at multiple scales.
- Images are represented by L_1-normalized 2-level spatial pyramids.
- For SVM, pyramid χ^2 kernel is used.



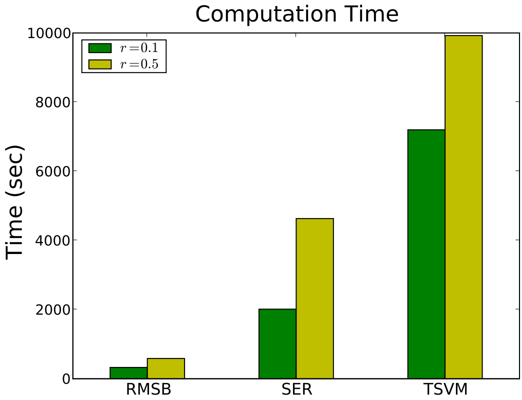
PASCAL 2006 Object Categorization Dataset



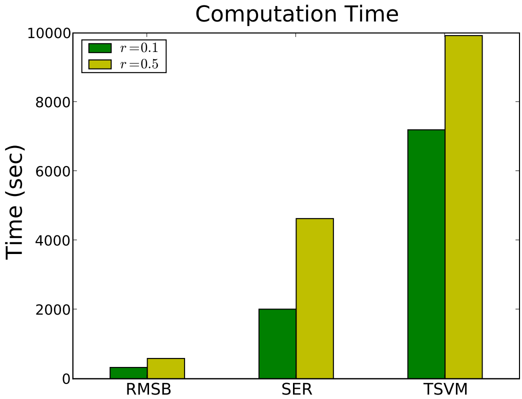
PASCAL 2006 Object Categorization Dataset



PASCAL 2006 Object Categorization Dataset



PASCAL 2006 Object Categorization Dataset



With our current GPU implementation of random forest, once can get a 10 to 20 times speed up here. An additional 5 times speed up can be achieved by reducing the iterations to 2000.

Conclusions: Release version 1.0

- We proposed a multi-class semi-supervised boosting method based on margin maximizing and cluster prior regularizations.
- By directly addressing the multi-class problem and using efficient base learners, such as random forests, we showed that our algorithm not only out-performs other supervised and semi-supervised methods, but also achieves a high level of computational efficiency.
- Additionally, our method provides a mean to incorporate other knowledge sources, such as label priors, knowledge transfer priors, or human knowledge.

Graz University of Technology

Amir Saffari, Christian Leistner, Horst Bischof

DAS-Forests

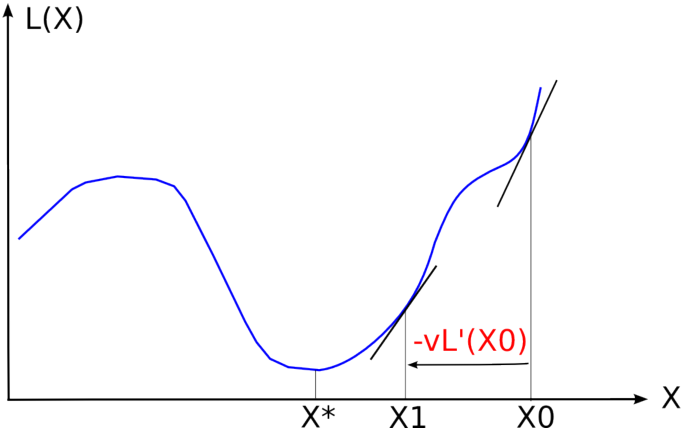
Semi-Supervised Random Forests, ICCV 2009.

Hope to see many of you at Kyoto.

TU Graz
Graz University of Technology

Learning with Functional Gradient Descent

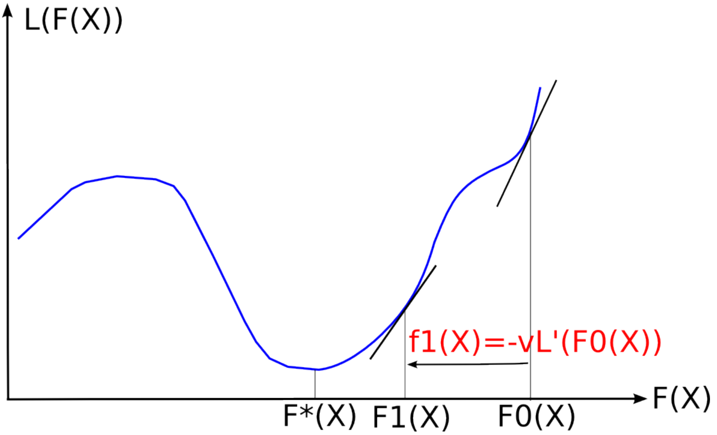
$$X^* = X_0 - \nu \sum_{t=1}^{T} L'(X_{t-1})$$



Friedman et al., Annals of Applied Statistics, 2000

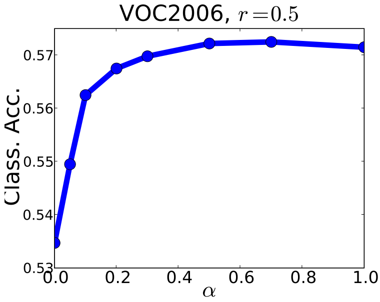
Learning with Functional Gradient Descent

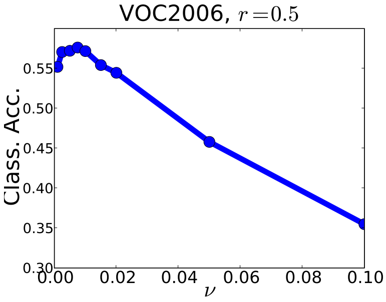
$$F^*(\mathbf{x}) = F_0(\mathbf{x}) - \nu \sum_{t=1}^{T} \frac{\partial L}{\partial F}\Big|_{(F_{t-1}(\mathbf{x}))}$$



Friedman et al., Annals of Applied Statistics, 2000

PASCAL 2006 Object Categorization Dataset





Graz University of Technology

Experimental Settings

- RMSBoost is compared with: AdaBoost.ML, Kernel SVM, Multi-Switch TSVM, SERBoost, RMBoost.
- Base learners are tiny extremely randomized forests, each consisting of 10 trees.

Graz University of Technology

Experimental Settings

- RMSBoost is compared with: AdaBoost.ML, Kernel SVM, Multi-Switch TSVM, SERBoost, RMBoost.
- Base learners are tiny extremely randomized forests, each consisting of 10 trees.
- Boosting iterations set to be 10000.

Graz University of Technology

Experimental Settings

- RMSBoost is compared with: AdaBoost.ML, Kernel SVM, Multi-Switch TSVM, SERBoost, RMBoost.
- Base learners are tiny extremely randomized forests, each consisting of 10 trees.
- Boosting iterations set to be 10000.
- Parameters are selected via 10-fold cross-validation.

Graz University of Technology

Experimental Settings

- RMSBoost is compared with: AdaBoost.ML, Kernel SVM, Multi-Switch TSVM, SERBoost, RMBoost.
- Base learners are tiny extremely randomized forests, each consisting of 10 trees.
- Boosting iterations set to be 10000.
- Parameters are selected via 10-fold cross-validation.
- Results of hierarchical k-means is averaged 10 times to estimate the cluster priors.

Graz University of Technology

Experimental Settings

- RMSBoost is compared with: AdaBoost.ML, Kernel SVM, Multi-Switch TSVM, SERBoost, RMBoost.
- Base learners are tiny extremely randomized forests, each consisting of 10 trees.
- Boosting iterations set to be 10000.
- Parameters are selected via 10-fold cross-validation.
- Results of hierarchical k-means is averaged 10 times to estimate the cluster priors.
- For binary classification methods, we used a 1-vs-all strategy.

TU Graz
Graz University of Technology

Experimental Settings

- RMSBoost is compared with: AdaBoost.ML, Kernel SVM, Multi-Switch TSVM, SERBoost, RMBoost.
- Base learners are tiny extremely randomized forests, each consisting of 10 trees.
- Boosting iterations set to be 10000.
- Parameters are selected via 10-fold cross-validation.
- Results of hierarchical k-means is averaged 10 times to estimate the cluster priors.
- For binary classification methods, we used a 1-vs-all strategy.
- All results reported are average of 10 independent runs.

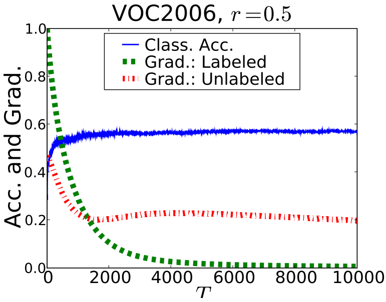
Graz University of Technology

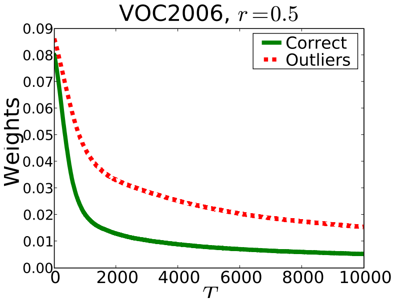
Experimental Settings

- RMSBoost is compared with: AdaBoost.ML, Kernel SVM, Multi-Switch TSVM, SERBoost, RMBoost.
- Base learners are tiny extremely randomized forests, each consisting of 10 trees.
- Boosting iterations set to be 10000.
- Parameters are selected via 10-fold cross-validation.
- Results of hierarchical k-means is averaged 10 times to estimate the cluster priors.
- For binary classification methods, we used a 1-vs-all strategy.
- All results reported are average of 10 independent runs.
- All boosting and RF methods are implemented in C++ and use ATLAS subroutines.

Graz University of Technology

PASCAL 2006 Object Categorization Dataset





Graz University of Technology

Exponential Loss

Example

The exponential loss $\ell(f(\mathbf{x})) = e^{-f(\mathbf{x})}$, is a Fisher-consistent loss, its estimated conditional probabilities can be written as

$$\hat{p}(y = i|\mathbf{x}) = \frac{e^{f_i(\mathbf{x})}}{\sum_{j=1}^{K} e^{f_j(\mathbf{x})}}, \tag{11}$$

which is a symmetric multiple logistic transformation.

The empirical risk is

$$\mathcal{L}(\mathbf{f}(\mathbf{x}), \mathcal{X}_l) = \sum_{(\mathbf{x},y)\in\mathcal{X}_l} e^{-f_y(\mathbf{x})}. \tag{12}$$

Zou et al., Annals of Applied Statistics, 2008

Graz University of Technology

Cluster Priors

Cluster Prior

$\forall \mathbf{x} \in \mathcal{X}_u, \forall i \in \{1, \cdots, K\} : p_p(y = i|\mathbf{x})$.

We use the Kullback-Leibler (KL) divergence to measure the deviation of the model w.r.t. cluster prior

$$\ell_c(\mathbf{f}(\mathbf{x})) = D(p_p \| \hat{p}) = -H(p_p) + H(p_p, \hat{p}). \tag{13}$$

Using symmetric multiple logistic transformation as the probabilistic estimates of the model

$$\ell_c(\mathbf{f}(\mathbf{x})) = -\mathbf{p}_p^T \mathbf{f}(\mathbf{x}) + \log \sum_{j=1}^{K} e^{f_j(\mathbf{x})}. \tag{14}$$

Graz University of Technology