MANCHESTER 1824

The University of Manchester

Regularized Boost for Semi-Supervised Learning

Ke Chen and Shihai Wang
School of Computer Science
The University of Manchester
Manchester M13 9PL, United Kingdom

Abstract

Semi-supervised inductive learning concerns how to learn a decision rule from a data set containing both labeled and unlabeled data. Several boosting algorithms have been extended to semi-supervised learning with various strategies. To our knowledge, however, none of them takes local smoothness constraints among data into account during ensemble learning. In this work, we introduce a local smoothness regularizer to semi-supervised boosting algorithms based on the universal optimization framework of margin cost functionals. Our regularizer is applicable to existing semi-supervised boosting algorithms to improve their generalization and speed up their training. Comparative results on synthetic, benchmark and real world tasks demonstrate the effectiveness of our local smoothness regularizer. We discuss relevant issues and relate our regularizer to previous work.

Background

❑ Pattern classification

- A ubiquitous task in the real world
- Supervised learning: demand labeled data for training
- Annotation of data: time consuming, expensive and difficult

❑ Semi-supervise learning

- A new paradigm for using unlabeled data to improve the performance of a classifier trained with fewer labeled data
- Objectives: find out a classification function
  - minimizes classification errors on labeled training data
  - compatible with the input distribution by inspecting their values on unlabeled data

Background

❑ Regularization in semi-supervised learning

- One of the most important factors for success in semi-supervised learning is properly exploiting unlabeled data.
- Data regularization: an effective way to exploit unlabelled data
  - Working on assumptions underlying data (Chapelle et al, 2006)
    - The smoothness assumption
    - The cluster assumption
    - The manifold assumption
  - Various regularization techniques
    - information-based regularization (Szummer & Jaakkola, 2003)
    - measure-based regularization (Bousquet et al, 2004)
    - graph-based regularization (Zhou et al, 2004; Zhu & Lafferty, 2005)
    - entropy-based regularization (Grandvalet & Begio, 2005)
    - manifold-based regularization (Belkin et al, 2005)

MANCHESTER 1824

The University of Manchester

Background

Semi-supervised boosting

- Boosting: an generic ensemble learning framework
  - Sequentially construct a linear combination of base learners that concentrate on difficult examples
  - In general, an optimization framework of margin functionals
- Extended to semi-supervised learning with different strategies
  - Semi-supervised MarginBoost (d'Alche-Buc et al, 2002)
  - ASSEMBLE (Bennett et al, 2002)
  - CoBoost (Collins & Singer, 1999)
  - Agreement Boost (Leskes, 2005)
- All semi-supervised boosting algorithms work in either a self-training or a co-training way.

Background

Semi-supervised boosting (cont.)

- The main problems of existing semi-supervised boosting
  - Without considering unlabelled data distribution
  - The use of confidence-based self-training or co-training only may cause a biased estimation on unlabelled data.

Objectives

- Tackling two aforementioned problems by exploiting the unlabelled data structure
- Introducing a universal local smoothness regularizer to existing semi-supervised boosting algorithms
- Demonstrating its effectiveness on various classification tasks

Semi-Supervised Boosting Learning

Problem statement

Given a training set $S = L \cup U$ $L = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_{|L|}, y_{|L|})\}$ $U = \{\mathbf{x}_{|L|+1}, \cdots, \mathbf{x}_{|L|+|U|}\}$ we wish to construct an ensemble learner $F(\mathbf{x}) = \sum_t w_t f_t(\mathbf{x})$ so that $P(F(\mathbf{x}) \neq y)$ small.

Ideas in semi-supervised boosting

- Introducing either pseudo-class labels or pseudo-margin for unlabeled data (Bennett et al, 2002; d'Alche-Buc et al, 2002)
- For binary classification ($y \in \{-1, 1\}$ the pseudo-class label is defined as $y = \text{sign}[F(\mathbf{x})]$
- The corresponding pseudo-margin is $yF(\mathbf{x}) = |F(\mathbf{x})|$

Semi-Supervised Boosting Learning

Ideas in semi-supervised boosting (cont.)

- Within the universal optimization framework of margin cost functional, the semi-supervised learning is to find F so that the following cost of functional is minimized.

$$\mathcal{C}(F) = \sum_{\mathbf{x}_i \in L} \alpha_i C[y_i F(\mathbf{x}_i)] + \sum_{\mathbf{x}_i \in U} \alpha_i C[|F(\mathbf{x}_i)|]$$

where $C : \mathbb{R} \to \mathbb{R}$'s a non-negative and monotonically decreasing cost function and the weight $\alpha_i \in \mathbb{R}^+$

- In each boosting round, find a base learner $\dot{f}(\mathbf{x})$maximizing

$$-\langle \nabla \mathcal{C}(F), f \rangle = \sum_{i:f(\mathbf{x}_i) \neq y_i} \alpha_i C'[y_i F(\mathbf{x}_i)] - \sum_{i:f(\mathbf{x}_i) = y_i} \alpha_i C'[y_i F(\mathbf{x}_i)]$$ or minimizing

$$\sum_{i:f(\mathbf{x}_i) \neq y_i} D(i) - \sum_{i:f(\mathbf{x}_i) = y_i} D(i) = 2 \underbrace{\sum_{i:f(\mathbf{x}_i) \neq y_i} D(i)}_{\text{misclassification errors}} - 1, \quad D(i) = \frac{\alpha_i C'[y_i F(\mathbf{x}_i)]}{\sum_{j \in S} \alpha_j C'[y_j F(\mathbf{x}_j)]}$$

Semi-Supervised Boosting Learning

Co-training based semi-supervised boosting

- Applying the above semi-supervised boosting procedure to each view of data to build up a component ensemble learner.
- Pseudo-class labels of unlabeled examples for a specific view is determined by other ensemble learners on different views.
- Agreement Boost (Leskes, 2005)
  - Co-training cost functional

$$\mathcal{C}(F^1, \cdots, F^J) = \sum_{j=1}^{J} \sum_{\mathbf{x}_i \in L} C[y_i F^j(\mathbf{x}_i)] + \eta \sum_{\mathbf{x}_i \in U} C[-V(\mathbf{x}_i)]$$

  - Disagreement of ensemble learners for an unlabeled example,

$$V(\mathbf{x}_i) = \frac{1}{J}\sum_{j=1}^{J}[F^j(\mathbf{x}_i)]^2 - [\frac{1}{J}\sum_{j=1}^{J}F^j(\mathbf{x}_i)]^2, \quad y_i = \text{sign}[\frac{1}{J}\sum_{j=1}^{J}F^j(\mathbf{x}_i) - F^j(\mathbf{x}_i)]$$

  - The use of pseudo-class labels leads to a proper base learner $\hat{f}^j(\mathbf{x})$ to be added to $F^j(\mathbf{x})$ for the jth view.

Regularized Boost

- Marginal cost functional with regularization
  - Introducing a regularization term to the margin cost functional

$$\mathcal{T}(F, f) = -\langle \nabla \mathcal{C}(F), f \rangle - \sum_{i:\mathbf{x}_i \in S} \beta_i R(i)$$

where $\beta_i \in \mathbb{R}^+$ is a weight determined by input data distribution and the local smoothness measure is defined as

$$R(i) = \sum_{j:\mathbf{x}_j \in S, j \neq i} W_{ij} \tilde{C}(-I_{ij}) \qquad W_{ij} = \exp(-||\mathbf{x}_i - \mathbf{x}_j||^2 / 2\sigma^2)$$

$I_{ij} = |y_i - y_j|$ is a class label compatibility function and y_i and y_j either true or pseudo-class labels of training examples.

  - Now finding a base leaner $f(\mathbf{x})$ needs to maximize

$$\mathcal{T}(F, f) = \sum_{i:f(\mathbf{x}_i) \neq y_i} \alpha_i C'[y_i F(\mathbf{x}_i)] - \sum_{i:f(\mathbf{x}_i) = y_i} \alpha_i C'[y_i F(\mathbf{x}_i)] - \sum_{i:\mathbf{x}_i \in S} \beta_i R(i).$$

Regularized Boost

Empirical data distribution

- The objective function yields a new empirical data distribution

$$\tilde{D}(i) = \frac{\alpha_i C'[y_i F(\mathbf{x}_i)] - \beta_i R(i)}{\sum_{j:\mathbf{x}_j \in S} \{\alpha_j C'[y_j F(\mathbf{x}_j)] - \beta_j R(j)\}}, \quad 1 \le i \le |L| + |U|.$$

Optimization problem

- Applying the empirical data distribution to $\mathcal{T}(F, f)$
- Finding a base learner $f(\mathbf{x})$ a maximizing $\mathcal{T}(F, f)$ equivalent to minimizing

$$2 \underbrace{\sum_{i:f(\mathbf{x}_i) \neq y_i} \tilde{D}(i)}_{\text{misclassification errors}} + 2 \underbrace{\sum_{i:f(\mathbf{x}_i) = y_i} \frac{-\beta_i R(i)}{\sum_{j:\mathbf{x}_j \in S} \{\alpha_j C'[y_j F(\mathbf{x}_j)] - \beta_j R(j)\}}}_{\text{local class label incompatibility}} - 1.$$

Regularized Boost

1. Initialization
   - 1.1 Input training set $S_0 = L \cup U$. For $i \in S_0$, assign α_i and estimate β_i. Further set a maximum iteration number, T, for boosting and initialize $F_0 = 0$.
   - 1.2 Set $\tilde{D}_0(i) = \zeta/|L|$ for $i \in L$ and $\tilde{D}_0(i) = (1-\zeta)/|U|$ for $i \in U$ where $0 \ll \zeta \leq 1$, and choose a base learner $\mathcal{L}(\cdot,\cdot,\cdot)$ as well as cost functions, $C(\cdot)$ and $\tilde{C}(\cdot)$.
   - 1.3 Assign the pseudo-class label y_i for $i \in U$ with class labels of its K nearest neighbors in L. $Y_U = \{y_i | i \in U\}$ is a collective notation of pseudo-class labels.
2. Repeat steps 2.1-2.4 for $t = 1, \cdots, T$
   - 2.1 $f_t = \mathcal{L}(S_{t-1}, Y_U, \tilde{D}_{t-1})$, and hence achieve $\hat{y}_i^{(t)} = f_t(\mathbf{x}_i)$ for $i \in L \cup U$.
   - 2.2 Calculate $\epsilon_t = \sum_{i:\hat{y}_i^{(t)} \neq y_i} \tilde{D}_{t-1}(i) + \sum_{i:\hat{y}_i^{(t)} = y_i} \frac{-\beta_i R_{t-1}(i)}{\sum_{j \in L \cup U} \left\{ \alpha_j C'[y_j F_{t-1}(\mathbf{x}_j)] - \beta_j R_{t-1}(j) \right\}}$.
   - 2.3 If $\epsilon_t > 1/2$, stop training and return the ensemble learner, F_{t-1}. Otherwise, $w_t = \frac{1}{2} \log\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$ so that $F_t = F_{t-1} + w_t f_t$. Reset $y_i = F_t(\mathbf{x}_i)$ for $i \in U$.
   - 2.4 Update $\tilde{D}_t(i) = \frac{\alpha_i C'[y_i F_t(\mathbf{x}_i)] - \beta_i R_t(i)}{\sum_{j \in L \cup U} \left\{ \alpha_j C'[y_j F_t(\mathbf{x}_j)] - \beta_j R_t(j) \right\}}$ for $i \in L \cup U$ to obtain a new training set by re-sampling; i.e., $S_t = \text{Sample}(L \cup U, Y_U, \tilde{D}_t)$.
3. Return the ensemble learner, F_T.

MANCHESTER 1824

The University of Manchester

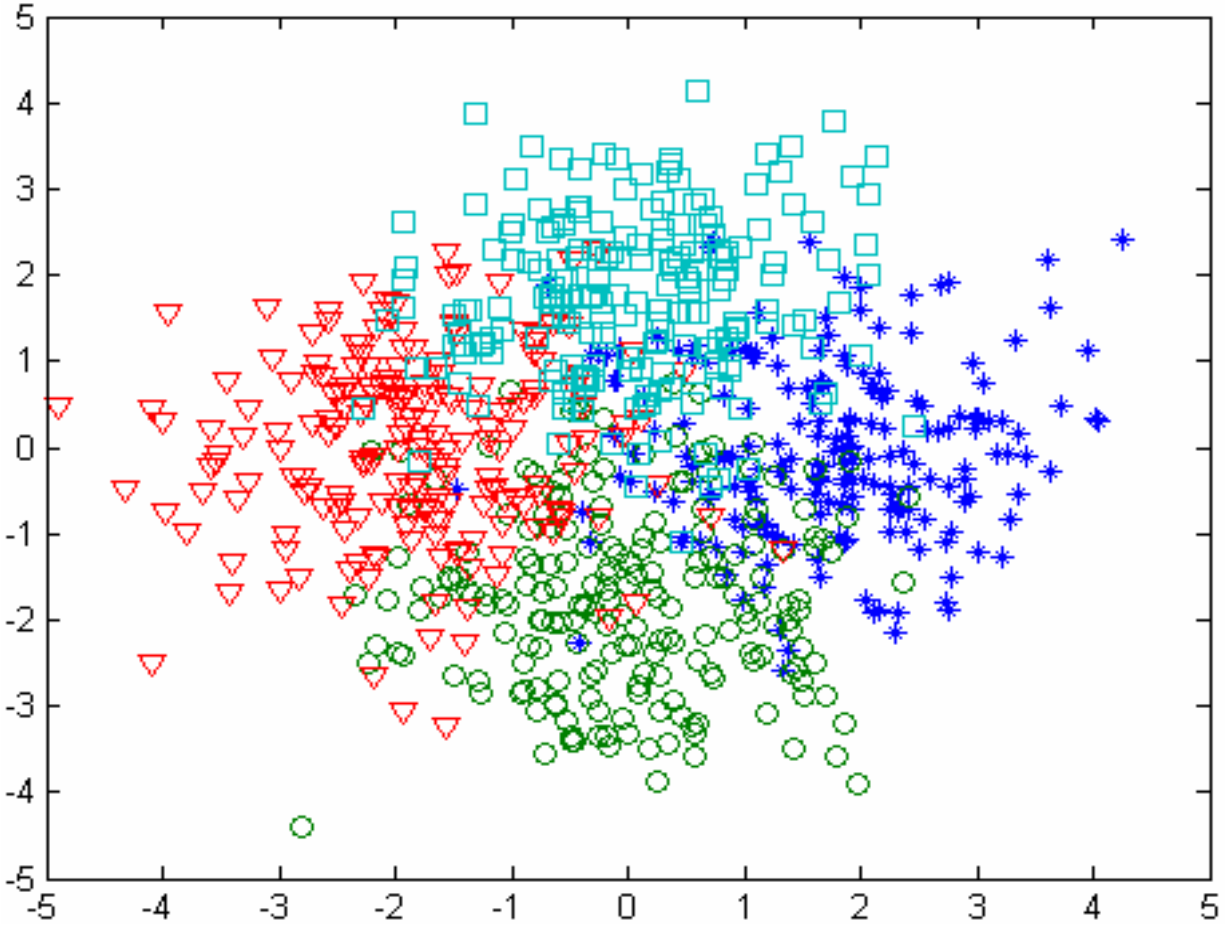
Experiment

Experimental methodology

- Our regularizer is applied to ASSEMBLE (Bennett et al, 2002), a winning algorithm in NIPS'01 Unlabeled Data Competition, and Agreement Boost (Leskes, 2005) to train its components.
- Algorithm setting
  - Cost functional: $C(\gamma) = e^{-\gamma}$ $\tilde{C}(\gamma) = C(\gamma) - 1$
  - Parameters: $\alpha_i = 1$ $\beta_i = \frac{1}{2}$, $T=100$
- Training vs. testing data
  - Data set randomly divided into training (80%) and testing (20%) data
  - Training set divided into labeled (L) and unlabeled (U) data with ratio 1:4
  - For each task, ten trials done for the above training vs. testing data setting
- Base learners
  - K Nearest Neighbors (KNN) and Multi-Layered Perceptron (MLP)
- Comparison with the original versions without regularization and also Adaboost trained on labeled data only

Experiment

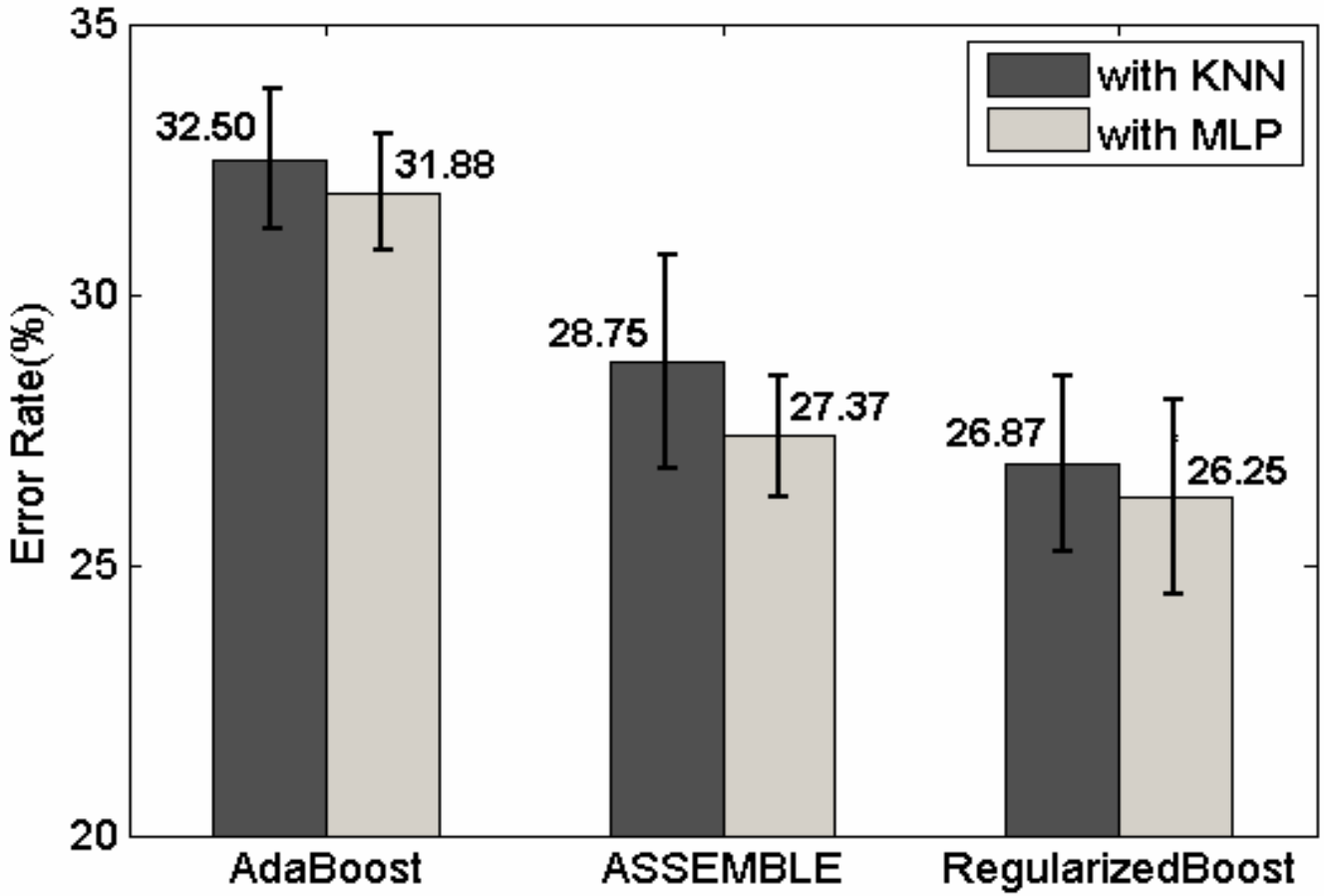
- Synthetic data
  - 4-category: 200 examples/category



Experiment

Synthetic data (cont.)

– Results of different algorithms with two base learners



Experiment

Results of ASSEMBLE

- Five UCI benchmark tasks of different classes
- OTIDIGITS data set
  - Pre-split into training and testing sets by the data collector
  - The training set randomly divided into labeled and unlabeled with a ratio 1:4
  - Comparison: 2.2% error rate by *3*-NN with all the labeled data
- Error rate: (mean±dev)%

Data Set	KNN			MLP		
	AdaBoost	ASSEMBLE	RegBoost	AdaBoost	ASSEMBLE	RegBoost
BUPA	37.7±3.4	36.1±3.0	34.9±3.1	35.1±1.1	31.2±6.7	28.8±5.6
WDBC	8.3±1.9	4.1±1.0	3.7±2.0	9.7±2.0	3.5±0.9	3.2±0.8
BSWD	22.2±0.9	18.7±0.4	17.4±0.9	16.8±2.8	14.4±2.4	13.6±2.6
CAR	31.3±1.2	24.4±0.7	23.2 ±1.1	30.6±3.0	20.5±0.9	17.7±1.1
OTIDIGITS	4.9±0.1	3.1±0.5	2.7±0.7	6.3±0.2	5.2±0.2	5.0±0.2

Experiment

❑ Results of Agreement Boost

– Five UCI benchmark tasks of binary classes used since the agreement boost (Leskes, 2005) can cope with a binary classification problem only.

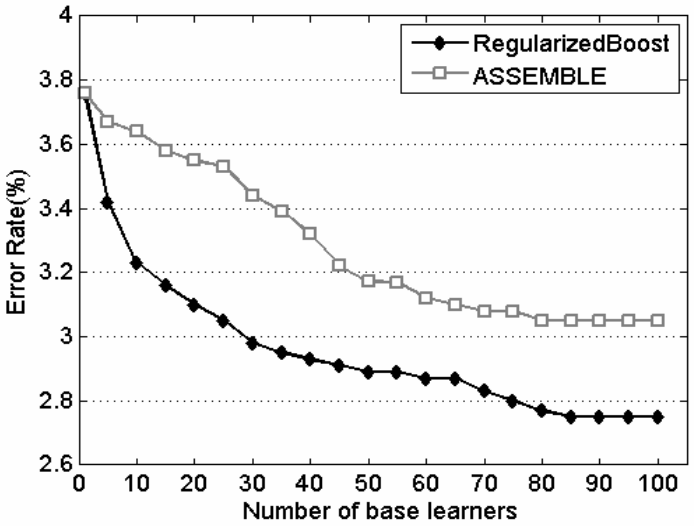
– Error rate: (mean±dev)%

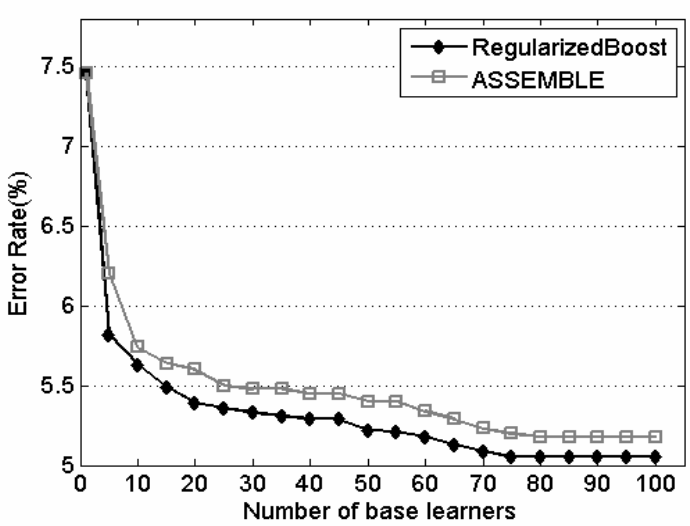
Data Set	AdaBoost-KNN	AdaBoost-MLP	AgreementBoost	RegBoost
BUPA	37.7±3.4	35.1±1.1	30.4±7.5	28.9±5.8
WDBC	8.3±1.9	9.7±2.0	3.3±0.7	3.0±0.8
VOTE	9.0±1.5	10.6±0.5	4.4±0.8	2.8±0.6
AUSTRALIAN	37.7±1.2	21.0±3.4	16.7±2.1	15.2±2.8
KR-vs-KP	15.6±0.7	7.1±0.2	6.3±1.3	5.2±1.6

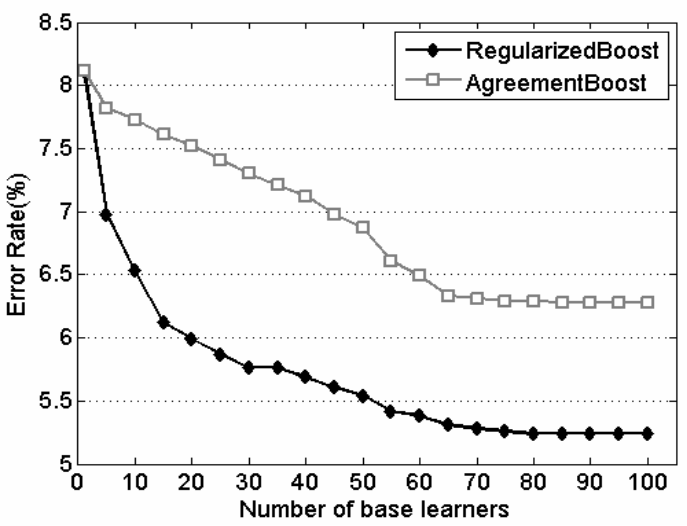
Experiment

Behaviors of Regularized Boost

- Stopping at different boosting rounds to achieve testing results
- Display the averaging error rates on ten trials on two UCI tasks
  - OPTDIGIS: ASSEMBLE with KNN or MLP
  - KR-vs-KP: Agreement Boost combining KNN and MLP





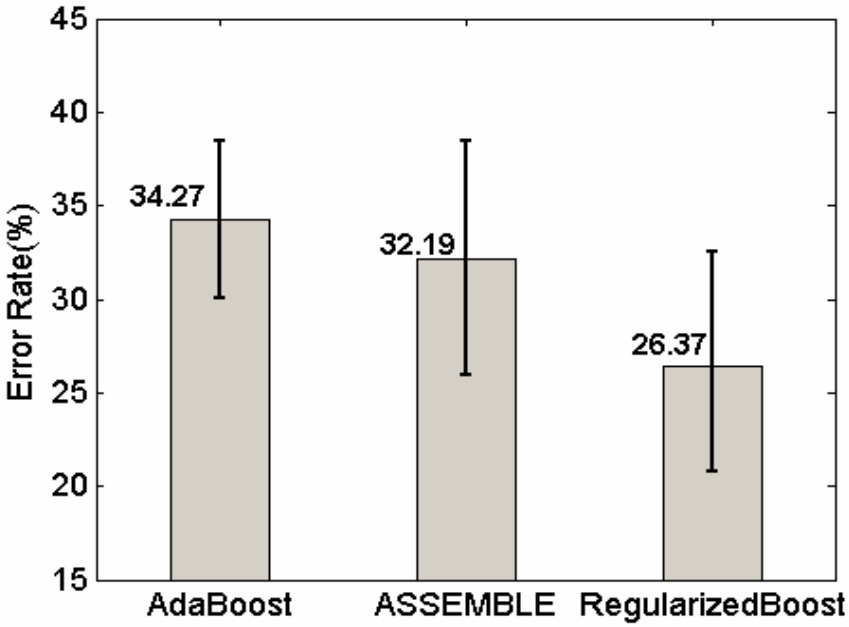


Experiment

- Facial expression recognition
  - JAFFE database
    - 213 facial images of 10 female expressers
    - Seven universal emotional states
  - Experimental setting
    - 20% images (balanced to seven classes) as testing data
    - Training set randomly split into labeled and unlabeled data of equal size
    - Feature extraction: ICA + PCA (40-D representation)
    - MLP of a single hidden layer with 30 neurons
    - Ten trials done with the about experimental methods

Experiment

- Facial expression recognition (cont.)



– Comparison: 31.5% achieved by a supervised convolution neural network with 70% images are used as training data (Fasel, 2002)

Discussion

Further issues of Regularized Boost

- The regularization is applied to each training point via a parameter which works β_i n the smoothness and the cluster assumptions by setting . $\beta_i = \lambda P(\mathbf{x})$
- We now use affinity metric system to measure the proximity, which can be extended to work on the manifold assumption by incorporating the manifold information.
- The local smoothness plays an important role in re-sampling all training data including both labeled and unlabeled data.
  - Our empirical data distribution encodes both misclassification and non-smoothing information.
  - Both labeled and unlabeled points are more likely to choose when located in a non-smoothing region
  - Potential problem: sensitive to noisy labeled data

Discussion

❑ Relation to Boosting on Manifolds (Kegl & Wang, 2004)

- Kegl & Wang (2004) proposed a boosting on manifolds algorithm via adaptive regularization of base classifiers.
- Their idea is using the graph Laplacian regularizer to select base learners based on decision boundaries and data structural information, which is identical to our objective.
- Their idea is simply applicable to the marginal Adaboost algorithm with adaptively adjusting the edge offset for a weight decay for both supervised and semi-supervised learning.
- In contrast, we encode the smoothness constraints in empirical data distribution, a simpler implementation of regularization for any boosting algorithms, especially for semi-supervised learning.

Discussion

❑ Relation to graph-based regularization (Zhou et al., 2004)

- In general, a graph-based semi-supervised learning method tends to find a function to simultaneously satisfy two constraints:
  - Close to given labels on the labeled nodes
  - Smooth on the whole graph
- Zhou et al (2004) proposed a regularization framework with a cost function to gain both local and global consistency:

$$\mathcal{Q}(F) = \frac{1}{2}\left(\mu \underbrace{\sum_i ||F_i - Y_i||^2}_{\text{fitting constraint}} + \underbrace{\sum_i \sum_j W_{ij} ||\frac{1}{\sqrt{D_{ii}}} F_i - \frac{1}{\sqrt{D_{jj}}} F_j||^2}_{\text{smoothness constraint}}\right)$$

  which is consistent with ours to find an optimal base leaner.
- A graph-based algorithm is an iterative label propagation process where the regularization directly responsible for label modification.
- Our regularized boost is an iterative process that runs a base learner on various empirical data distributions where our regularization simply plays a role in determining such distributions.

Conclusion

- We have proposed a regularized boost approach for semi-supervised learning.
- Our approach is applicable to any existing semi-supervised boosting algorithms by introducing a regularization term to the optimization framework of margin cost functional.
- Experimental results on different type of classification tasks demonstrate its effectiveness and efficiency.
- Our ongoing research is tackling the “noisy” labelling problem that could result in overfitting as well as working on a theoretical analysis to explain the behaviors of our algorithm formally.