Information Theoretic Regularization for Semi-Supervised Boosting

Lei Zheng[1], Shaojun Wang[1], Yan Liu[1], Chi-Hoon Lee[2]
[1]Kno.e.sis Center
Wright State University
[2]Yahoo! Lab
lei.zheng@wright.edu
Presented by Chris Ding, University of Texas at Arlington

Outline

- Introduction
- Boosting As An Optimization Method
- Generic Semi-supervised Boosting Algorithm
- Information Theoretic Regularization Approach
- Experiment results and conclusions

Boosting and Semi-supervised Learning

- Boosting
  - supervised learning methods
  - AdaBoost algorithm (Freund and Schapire (1997)
  - Various variants of AdaBoost algorithm
- Supervised learning: $D_l=(x_1,y_1),\ldots,(x_N, y_N)$
- Unsupervised learning: $D_u=(x_{N+1}, x_{N+2},\ldots,x_M)$
- Semi-supervised learning
  - Use both D_l and D_u
  - Supervised learning + Additional unlabeled data
  - Unsupervised learning + Additional labeled data

Semi-supervised Methods

- EM with generative model
- Self learning: classification EM algorithm (in statistics); bootstrapping (in NLP)
- Co-training
- Information regularization: mutual information and entropy regularization
- Graph-based transductive method: undirected graph Laplacian or directed graph Laplacian

Information Regularization for Semi-Supervised Boosting

$$\text{Semi-supervised Boosting} \begin{cases} \text{Entropy Regularization} \\ \\ \text{Mutual Information Regularization} \end{cases}$$

- **Motivation**

  Minimizing conditional entropy or mutual information over unlabeled data encourages the algorithm to find putative labelings for the unlabeled data that are mutually reinforcing with the supervised labels.

Outline

- Introduction
- Boosting As An Optimization Method
- Generic Semi-supervised Boosting Algorithm
- Information Theoretic Regularization Approach
- Experiment results and conclusions

Two Basic Approaches

- Boosting
  - Maximum Entropy Approach

    Described as a greedy feature induction algorithm that incrementally builds random fields to solve the maxent problem. The greediness of the algorithm arises in steps that select the most informative feature.

  - Greedy Function Optimization

    Statistical models are typically additive expansions in a set of basis functions and are fitted by minimizing a loss function averaged over the training data.

we adopt the Greedy Function Optimization approach

Outline

- Introduction
- Boosting As An Optimization Method
- Generic Semi-supervised Boosting Algorithm
- Information Theoretic Regularization Approach
- Experiment results and conclusions

∈

Surrogoate Loss

To minimize the following Surrogate Loss over 0/1 loss

$$J(h_t) = \underbrace{\sum_{i=1}^{N} L_l(y_i h_t(x_i))}_{\text{Labeled Data}} + \underbrace{\gamma}_{\text{trade-off parameter}} \underbrace{\sum_{i=N+1}^{M} L_u(h_t(x_i))}_{\text{Unlabeled Data}}$$

suppose that we have already included t-1 component classiers

$$h_{t-1}(x) = \lambda_1 h(x;\theta_1) + \dots + \lambda_{t-1} h(x;\theta_{t-1})$$

To add another $h(x;\theta)$:

$$J(h_t) = \sum_{i=1}^{N} L_l(y_i h_{t-1}(x_i) + y_i \lambda h(x_i;\theta)) + \gamma \sum_{i=N+1}^{M} L_u(h_t(x_i) + \lambda h(x_i;\theta))$$

$\lambda,\ \theta$: two parameters to optimize

Minimization Surrogoate Loss

Implement the optimization approximately in two steps

1. Find the new parameters θ so as to maximize its potential in reducing the surrogate loss. More precisely, set θ so as to minimize the derivative

$$\frac{d}{d\lambda} J(\lambda,\theta)|_{\lambda=0} = \sum_{i=1}^{N} dL_l(y_i h_{t-1}(x_i)) y_i h(x_i;\theta) + \gamma \sum_{i=N+1}^{M} \sum_{y} dL_u(y h_{t-1}(x_i)) y h(x_i;\theta)$$

2. After find $\hat{\theta}$, solve the minimization problem for λ_t over the following objective function:

$$J(\lambda,\hat{\theta}_t) = \sum_{i=1}^{N} L_l(y_i h_{t-1}(x_i) + y_i \lambda h(x_i;\hat{\theta}_t)) + \gamma \sum_{i=N+1}^{M} L_u(h_t(x_i) + \lambda h(x_i;\hat{\theta}_t))$$

This can be done by one-dimensional numerical line search

Outline

- Introduction
- Boosting As An Optimization Method
- Generic Semi-supervised Boosting Algorithm
- Information Theoretic Regularization Approach
- Experiment results and conclusions

Entropy and Mutual Information Regularization (Binary Classification)

- $y \in \{-1, 1\}$
- Normalized log-linear models: $p(y \mid x) = \frac{e^{(-yh(x))}}{\sum_y e^{(-yh(x))}}$
- Logistic loss for labeled data:

$$L_l(y_i h_t(x_i)) = -\log p(y_i \mid x_i) = \log(1 + e^{(-y_i h_t(x_i))})$$

- For unlabeled data:
  - Entropy regularization

$$L_u(h_t(x_i)) = \sum_y L_u(yh_t(x_i)) = H(p(y \mid x_i))$$

  - Mutual Information regularization

$$L_u(h_t(x_i)) = \sum_y L_u(yh_t(x_i)) = H(p(y)) - H(p(y \mid x_i))$$

Multi-class Classification

- Recode the class label $y \in \mathcal{Y}=\{1,\ldots, K\}$ with a K-dimesional vector ***c***, with all entries equal to -1/(K-1) except a 1 in position k if y = k. (Zhu et al. 2005)
- The normalized log-linear model

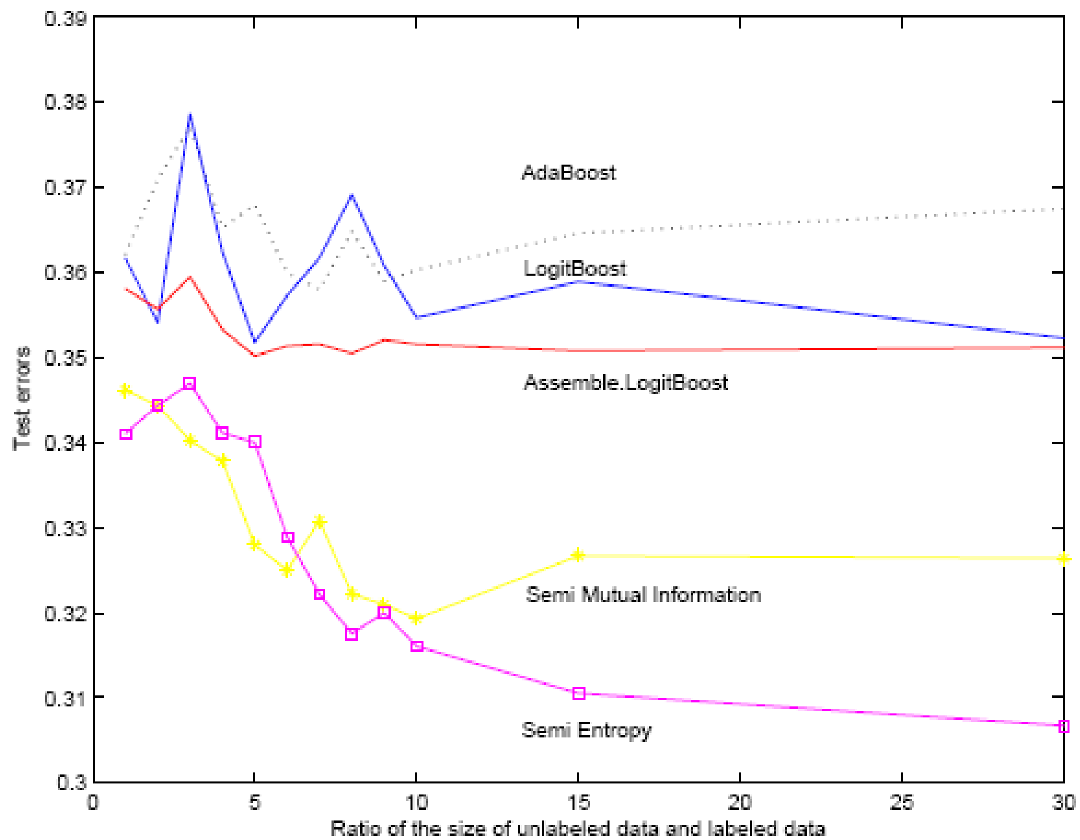
$$p(y \mid x) = \frac{e^{(-\frac{1}{K} C(y)^T h(x))}}{\sum_y e^{(-\frac{1}{K} C(y)^T h(x))}}$$

- The loss for labeled data and the loss for unlabeled data (mutual information and entropy) are simple math
- The process to minimize the loss is the same as we presented before.

Outline

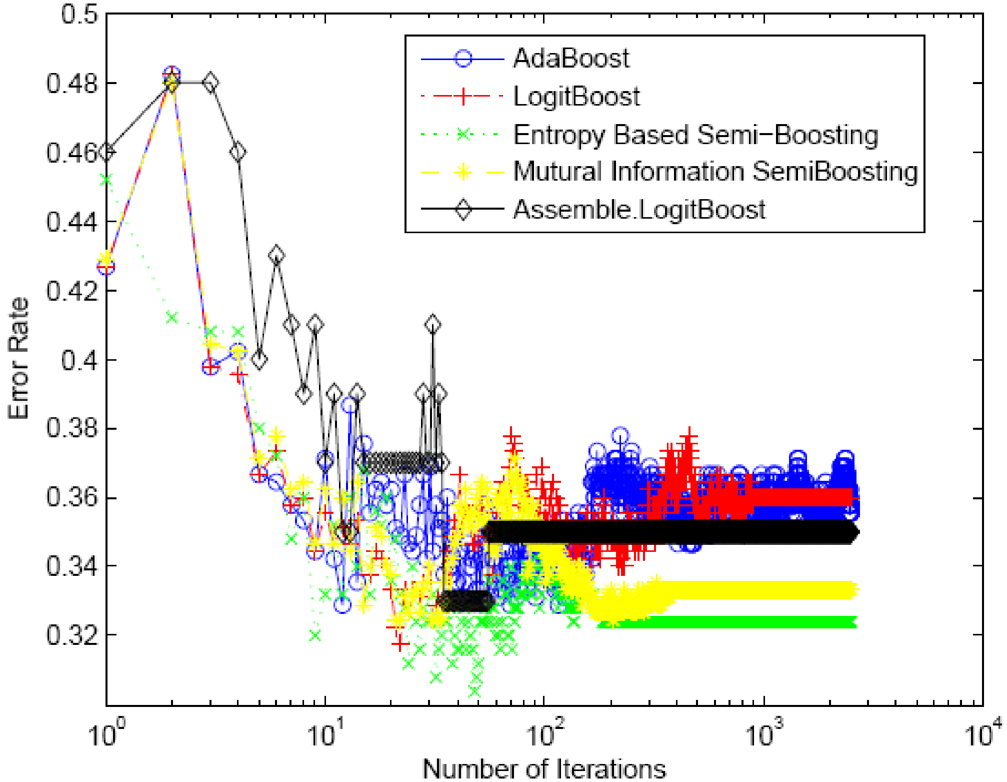
- Introduction
- Boosting As An Optimization Method
- Generic Semi-supervised Boosting Algorithm
- Information Theoretic Regularization Approach
- Experiment results and conclusions

Experimental 1: Synthetic Data



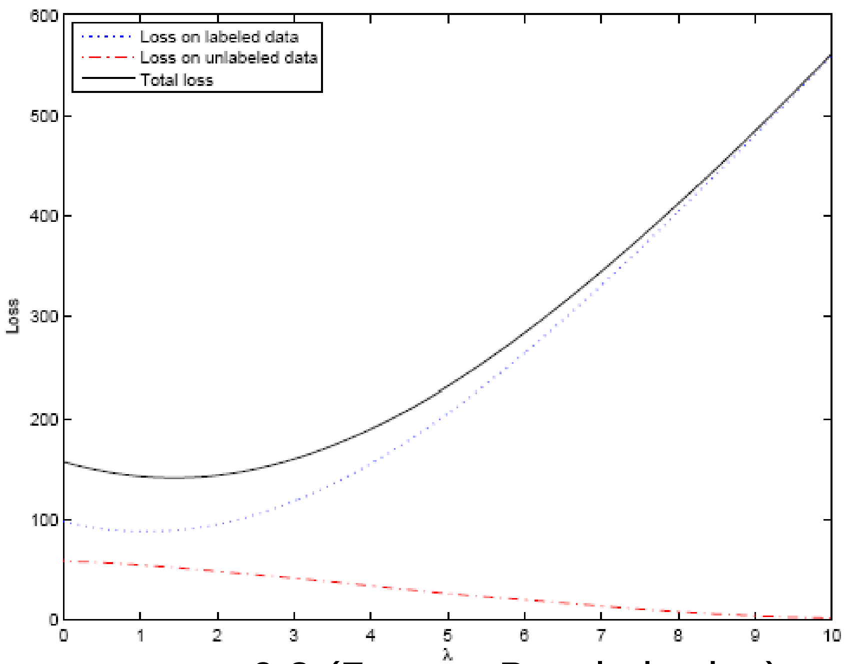
Test errors on data generated by two mixtures of 10-dimensional Gaussian distribution when we increase the size of unlabeled data.

Experiment 1

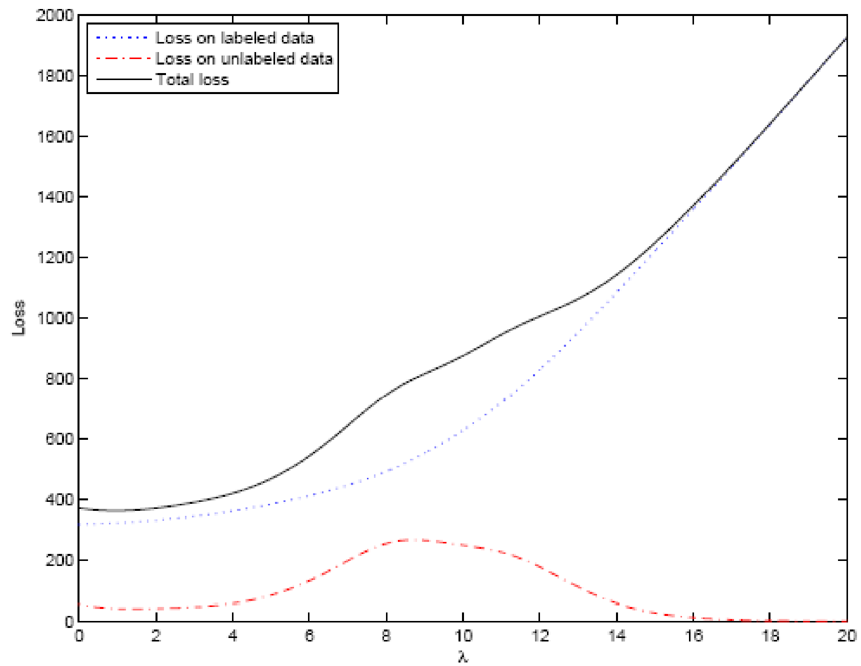


Test errors vary at each iteration with maximum iteration being 2500 where the ratio of unlabeled data and labeled data is set to 5.

Experiment 1



γ= 0.2 (Entropy Regularization)



γ = 1 (Entropy Regularization)

The loss function on labeled data is convex and the loss function on unlabeled data is non-convex. When the regularization parameter is small, total loss function to be convex; When the regularization parameter is large, the total loss function is non-convex.

Experiment 2: Benchmark Data

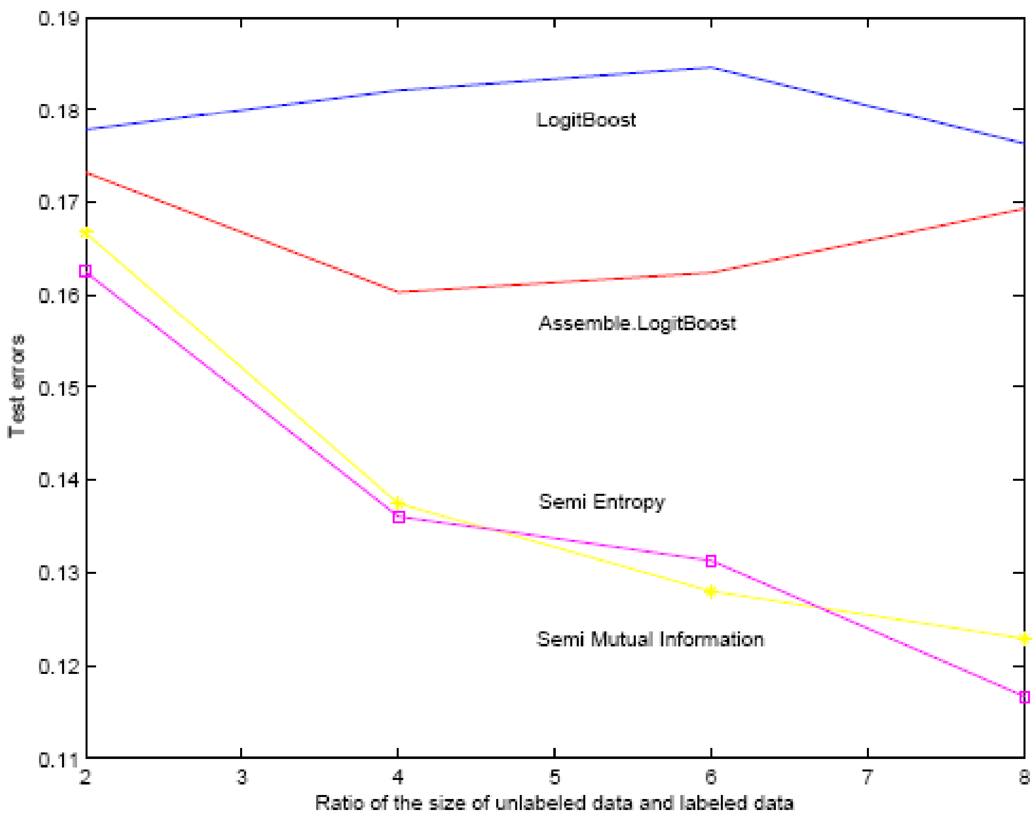
- UCI Machine Learning Repository
  - 15% as labeled data and 85% as unlabeled data
  - unlabeled data are used as the test data

Data	Logit	Assemble	MI	Entropy
Bala	27.43(1.52)	25.76(1.47)	24.80(1.72)	24.10(2.02)
Pima	22.50(2.52)	20.87(3.47)	20.44(3.75)	19.87(3.03)
Wins	5.14(0.74)	4.15(1.12)	2.92(0.77)	3.77(1.07)
BUPA	37.24(5.59)	36.17(3.40)	29.84(3.79)	31.77(2.31)

Error rates (%) on four benchmark UCI data sets

Experiment 3: Real Data

- Real EEG data to model human work load (2-class case)



Test errors on EEG data when we increase the size of unlabeled data

Experiment 3

- 3-class case of EEG
  - the number of labeled data is 30
  - the number of unlabeled data is 70
  - Unlabeled data are still the test data
- Error rate
  - LogitBoost: 32.94% (2.47)
  - Assemble. LogitBoost: 31.01% (1.97)
  - Entropy semi-supervised boosting 29.43% (2.44)
  - Mutual information semi-supervised boosting 30.58% (2.51)

Conclusions

- Semi-supervised boosting learning
  - information theoretic terms are used to encode the information provided by unlabeled data and behave as data dependent priors.
- The combined loss functions are non-convex
  - simple sequential gradient descent optimization algorithms
  - test these algorithms on synthetic, benchmark and real world tasks.
- Impressively improve the performances of supervised boosting algorithms
- We are working on a formal analysis to give some theoretical justifications.

- Thank you!
- Questions?