Batch Mode Active Learning and Its Application to Medical Image Classification ICML 2006

S. Hoi, R. Jin, J. Zhu, M. Lyu
Presenter: Esther Wang

February 19, 2009

Table of contents

Active Learning/Pool-based Active Learning (1)

Method:

1. Choose example with highest classification uncertainty for manual labeling

Active Learning/Pool-based Active Learning (1)

Method:

1. Choose example with highest classification uncertainty for manual labeling
2. Retrain classification model with new labeled example

Active Learning/Pool-based Active Learning (1)

Method:

1. Choose example with highest classification uncertainty for manual labeling
2. Retrain classification model with new labeled example
3. Iterate until most examples can be classified with reasonable confidence

Active Learning/Pool-based Active Learning (1)

Method:

1. Choose example with highest classification uncertainty for manual labeling
2. Retrain classification model with new labeled example
3. Iterate until most examples can be classified with reasonable confidence

Wish list:

- **Minimum requirement:** Generalization error should $\longrightarrow 0$ asymptotically
- **Fallback guarantee:** Convergence rate of error of active learning "at least as good" as passive learning
- **Rate improvement:** Error of active learning decreases much faster than for passive learning.

Goal: Label as little data as possible to achieve the confidence

Active Learning/Pool-based Active Learning (2)

- How do we measure the classification uncertainty of the unlabeled examples?

Active Learning/Pool-based Active Learning (2)

- How do we measure the classification uncertainty of the unlabeled examples?
  - What is the disagreement among ensemble of classification models in predicting labels for test examples?

Active Learning/Pool-based Active Learning (2)

- How do we measure the classification uncertainty of the unlabeled examples?
  - What is the disagreement among ensemble of classification models in predicting labels for test examples?
  - How far are away are the examples from the classification boundary, i.e. classification margin?

Active Learning/Pool-based Active Learning (2)

- How do we measure the classification uncertainty of the unlabeled examples?
  - What is the disagreement among ensemble of classification models in predicting labels for test examples?
  - How far are away are the examples from the classification boundary, i.e. classification margin? SVM (Tong & Koller, 2000)

Active Learning/Pool-based Active Learning (2)

- How do we measure the classification uncertainty of the unlabeled examples?
  - What is the disagreement among ensemble of classification models in predicting labels for test examples?
  - How far are away are the examples from the classification boundary, i.e. classification margin? SVM (Tong & Koller, 2000)
- **Problem:** Only a single example is selected for manual labeling at each iteration

Active Learning/Pool-based Active Learning (2)

- How do we measure the classification uncertainty of the unlabeled examples?
  - What is the disagreement among ensemble of classification models in predicting labels for test examples?
  - How far are away are the examples from the classification boundary, i.e. classification margin? SVM (Tong & Koller, 2000)
- **Problem:** Only a single example is selected for manual labeling at each iteration
- **Solution:** Use batch mode active learning to select examples that are most informative

Active Learning/Pool-based Active Learning (2)

- How do we measure the classification uncertainty of the unlabeled examples?
  - What is the disagreement among ensemble of classification models in predicting labels for test examples?
  - How far are away are the examples from the classification boundary, i.e. classification margin? SVM (Tong & Koller, 2000)
- **Problem:** Only a single example is selected for manual labeling at each iteration
- **Solution:** Use batch mode active learning to select examples that are most informative

Applications in Medical Image Classification

- Active learning has applications in text categorization, computer vision & information retrieval

Applications in Medical Image Classification

- Active learning has applications in text categorization, computer vision & information retrieval
- Few image categorization studies are devoted to the medical domain

Applications in Medical Image Classification

- Active learning has applications in text categorization, computer vision & information retrieval
- Few image categorization studies are devoted to the medical domain
  - Hospitals manage several tera-bytes of medical image data/year
  - Categorization of medical images is very important! Especially in digital radiology such as computer-aided diagnosis or case-based reasoning (Lehmann et al., 2004)
  - Expensive to acquired labeled data!

Batch Mode Active Learning

- Choose the top k most uncertain examples?

Batch Mode Active Learning

- Choose the top k most uncertain examples?
  Examples could be strong correlated!

Batch Mode Active Learning

- Choose the top k most uncertain examples? Examples could be strong correlated!
- We want examples that are:
  - Informative to the classification model
  - Diverse

Batch Mode Active Learning

- Choose the top k most uncertain examples?
  Examples could be strong correlated!
- We want examples that are:
  - Informative to the classification model
  - Diverse
- Challenges:
  1. How do we measure the "goodness" of the selected examples?
  2. How do we solve the related optimization problem?

Batch Mode Active Learning

- Choose the top k most uncertain examples?
  Examples could be strong correlated!
- We want examples that are:
  - Informative to the classification model
  - Diverse
- Challenges:
  1. How do we measure the "goodness" of the selected examples?
  2. How do we solve the related optimization problem?

General Overview

We want to pick examples that are

1. Informative to the classification model
2. Diverse so that the information provided by individual examples does not overlap

General Overview

We want to pick examples that are

1. Informative to the classification model
2. Diverse so that the information provided by individual examples does not overlap

Methods:

1. Use Fisher information matrix as a measurement of model (logistic regression) uncertainty
2. Use kernel trick to extend the linear classification model to nonlinear classification
3. Use greedy algorithm that optimizes submodular set function $f(S)$

Logistic Regression (1)

- In **multiple regression analysis**, continuous outcome variable is a linear combination of a set of predictors and error

$$Y = \alpha + \beta_1 X_1 + \cdots + \beta_n X_n + \epsilon = \alpha + \sum_{i-1}^{n} \beta_i X_i + \epsilon \qquad (1)$$

Logistic Regression (1)

- In **multiple regression analysis**, continuous outcome variable is a linear combination of a set of predictors and error

$$Y = \alpha + \beta_1 X_1 + \cdots + \beta_n X_n + \epsilon = \alpha + \sum_{i-1}^{n} \beta_i X_i + \epsilon \qquad (1)$$

- In **logistic regression analysis**, Y is categorical, i.e. binary

$$log\left(\frac{P(Y=1 \mid X_1, \ldots X_n)}{1 - P(Y=1 \mid X_1, \ldots, X_n)}\right) = log\left(\frac{\pi}{1-\pi}\right) \qquad (2)$$

$$= \alpha + \beta_1 X_1 + \cdots + \beta_n X_n = \alpha + \sum_{i=1}^{n} \beta_i X_i \qquad (3)$$

$$P(x) = \frac{1}{1 + exp(-(\alpha + \beta^T * X))} \qquad (4)$$

Problem Formulation: Binary Classification Problem

- **Goal:** Predict label $y \in \{-1, 1\}$ for given data x, want to find distrbution paramter α s.t. the joint distribution is $p(x, y) = p(x, y \mid \alpha)$
- Use statistical methods to analyze effect of unlabeled data on efficiency of paramter estimation
- Semi-parametric model: $p(x, y \mid \alpha) = p(x)p(y \mid x, a)$
- Logistic model: $p(x, y \mid \alpha) = (1 + exp(-\alpha^T xy))^{-1} p(x)$
- Use MLE to determine regularized logistic regression model parameter: $\hat{\alpha} = \text{argmin}_\alpha E_n log(1 + exp(-\alpha^T xy) + \lambda\alpha^2$

Fisher Information Matrix

- Cramér-Rao lower-bound: for any unbiased estimator t_n of α based on n i.i.d. samples from $p(x, y \mid \alpha)$, the covariance of t_n satisfies:

$$cov(t_n) \geq \frac{1}{n} I(\alpha)^{-1} \tag{5}$$

where

$$I(\alpha) = -\int p(x, y \mid \alpha) \frac{\partial^2}{\partial \alpha^2} log\ p(x, y \mid \alpha) dxdy \tag{6}$$

is the Fisher information matrix

- MLE achieves this lower bound & is unbiased asymptotically, so the MLE is the asymptotically most efficient (unbiased) estimator (Zhang & Oles, 2000)
- Represents overall uncertainty of a classifier

Fisher Information Matrix

- $p(\mathbf{x})$: distr. of all unlabeled examples
- $q(\mathbf{x})$: distr. of unlabeled examples chosen for manual labeling
- α: parameters of the classification model
- $I_p(\alpha)$ & $I_q(\alpha)$: Fisher info. matrix of classification for $p(\mathbf{x})$ & $q(\mathbf{x})$
- Minimize

$$q^* = \arg\min_q \mathrm{tr}(I_q(\alpha)^{-1} I_p(\alpha)) \quad (7)$$

$$I_q(\alpha) = -\int q(\mathbf{x}) \sum_{u=\pm 1} p(y|\mathbf{x}) \frac{\partial^2}{\partial \alpha^2} \log p(y|\mathbf{x}) d\mathbf{x} \quad (8)$$

$$= \int \frac{1}{1+e^{\alpha^T x}} \frac{1}{1+e^{-\alpha^T x}} \mathbf{x}\mathbf{x}^T q(\mathbf{x}) d\mathbf{x} \quad (9)$$

Fisher Information Matrix for Logistic Regression Models

Estimate optimal distribution $q(\mathbf{x})$:

$$I_p(\hat{\alpha}) = \frac{1}{n}\sum_{x\in D}\pi(\mathbf{x})(1-\pi(\mathbf{x}))\mathbf{x}\mathbf{x}^T + \delta I_d \quad (10)$$

$$I_q(S,\hat{\alpha}) = \frac{1}{k}\sum_{x\in S}\pi(\mathbf{x})(1-\pi(\mathbf{x}))\mathbf{x}\mathbf{x}^T + \delta I_d \quad (11)$$

$D = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$: unlabeled data
$S = (\mathbf{x}_1^s, \mathbf{x}_2^s, \ldots, \mathbf{x}_k^s)$: subset of selected examples
$\hat{\alpha}$: classification model estimated from labeled examples
k: number of examples selected
$\pi(\mathbf{x}) = p(-|\mathbf{x}) = \frac{1}{1+exp(\hat{\alpha}^T\mathbf{x})}$
$\delta << 1$: smoothing parameter

Final Optimization Problem for Batch Mode Active Learning

$$S^* = \operatorname{argmin}_{S \subseteq D \wedge |S| = k} \operatorname{tr}(I_q(S, \hat{\alpha})^{-1} I_p(\alpha)) \tag{12}$$

Apply Result to the Nonlinear Classification Model (1)

- Rewrite logistic regression with kernel function $K(x', x)$ (Zhu & Hastie, 2001):

$$p(y \mid x) = \frac{1}{1 + exp(-yK(w, x))} \tag{13}$$

- Use Representer Theorem to rewrite $\phi(w)$:

$$\phi(w) = \sum_{x \in L} \theta(x)\phi(x) \tag{14}$$

$\theta(x)$: combination weight for labeled xamples x,
$L = ((y_1, x_1^L), \ldots, (y_m, x_m^L))$: set of labeled examples,
m: # labeled examples

Apply Result to the Nonlinear Classification Model (2)

- Rewrite $K(w,x)$ and $p(y \mid x)$:

$$K(w,x) = \sum_{x' \in L} \theta(x')K(x',x) \tag{15}$$

$$p(y|x) = \frac{1}{1 + exp(-y\sum_{x' \in L} \theta(x')K(x',x))} \tag{16}$$

- Let $(K(x_1^L, x), \ldots, K(x_m^L, x))$ be the representation for unlabeled example x and directly apply results of linear logistic regression model

Key Idea

Optimization problem:

$$S^* = \text{argmin}_{S \subseteq D \wedge |S|=k} \text{tr}(I_q(S, \hat{\alpha})^{-1} I_p(\alpha)) \tag{17}$$

Challenge: # of candidate sets for S is exponential in n

Key Idea

Optimization problem:

$$S^* = \text{argmin}_{S \subseteq D \wedge |S|=k} \text{tr}(I_q(S, \hat{\alpha})^{-1} I_p(\alpha)) \tag{17}$$

Challenge: $\#$ of candidate sets for S is exponential in n
Solution: Use a submodular function!

Submodular Approximation to the Optimization Problem

Theorem about submodular functions (Nemhauser et al., 1987):

- $max_{|S|=k} f(S)$
- Greedy algorithm guarantees performance $(1 - 1/e)f(S^*)$, where $S^* = \text{argmax}_{|S|=k} f(S)$ is the optimal set if $f(S)$ is:
  1. Nondecreasing submodular function
  2. $f(\emptyset) = 0$

Submodular Approximation to the Optimization Problem

Theorem about submodular functions (Nemhauser et al., 1987):

- $max_{|S|=k} f(S)$
- Greedy algorithm guarantees performance $(1 - 1/e)f(S^*)$, where $S^* = \text{argmax}_{|S|=k} f(S)$ is the optimal set if $f(S)$ is:
  1. Nondecreasing submodular function
  2. $f(\emptyset) = 0$
- ...a bunch of algebra later, the optimization problem simplifies to $max_{|S|=k \wedge S \subseteq D} f(S)$, where set function $f(S)$ is

$$f(S) = \frac{1}{\delta} \sum_{x \in D} \pi(x)(1 - \pi(x))$$
$$- \sum_{x \notin S} \frac{\pi(x)(1 - \pi(x))}{\delta + \sum_{x' \in S} \pi(x')(1 - \pi(x'))(x^T x')^2}$$

A Greedy Algorithm for $\text{argmax}_{x \notin S} f(S)$

- **Initialize** $S = \emptyset$
- **For** $i = 1, 2, \ldots, k$

  Compute $x^* = \text{argmax}_{x \notin S} f(S \cup x) - f(S)$

  Set $S = S \cup x^*$

Value of the subset found by the greedy algorithm is $\geq 1 - 1/e$ the value of the true optimal subset

Analysis of Difference Between $f(S \cup \mathbf{x})$ and $f(S)$

$$\overbrace{f(S \cup x)}^{A} - f(S) = \overbrace{g(x, S)}^{B} + \overbrace{\sum_{x' \notin (S \cup x)} g(x', S) g(x, S \cup x)(x^T x')^2}^{C}$$

$$g(x, S) = \frac{\pi(x)(1 - \pi(x))}{\underbrace{\delta + \sum_{x' \in S} \pi(x')(1 - \pi(x'))(x^T x')^2}_{D}}$$

(1) $A \propto \pi(x)(1 - \pi(x))$ Uncertain to current classification model

Analysis of Difference Between $f(S \cup \mathbf{x})$ and $f(S)$

$$\overbrace{f(S \cup x)}^{A} - f(S) = \overbrace{g(x, S)}^{B} + \overbrace{\sum_{x' \notin (S \cup x)} g(x', S) g(x, S \cup x)(x^T x')^2}^{C}$$

$$g(x, S) = \frac{\pi(x)(1 - \pi(x))}{\underbrace{\delta + \sum_{x' \in S} \pi(x')(1 - \pi(x'))(x^T x')^2}_{D}}$$

(1)	$A \propto \pi(x)(1 - \pi(x))$	Uncertain to current classification model
(2)	$B \propto \frac{1}{D}$	Dissimilar to other selected examples

Analysis of Difference Between $f(S \cup \mathbf{x})$ and $f(S)$

$$\overbrace{f(S \cup x)}^{A} - f(S) = \overbrace{g(x, S)}^{B} + \overbrace{\sum_{x' \notin (S \cup x)} g(x', S) g(x, S \cup x)(x^T x')^2}^{C}$$

$$g(x, S) = \frac{\pi(x)(1 - \pi(x))}{\underbrace{\delta + \sum_{x' \in S} \pi(x')(1 - \pi(x'))(x^T x')^2}_{D}}$$

(1)	$A \propto \pi(x)(1 - \pi(x))$	Uncertain to current classification model
(2)	$B \propto \frac{1}{D}$	Dissimilar to other selected examples
(3)	$C \propto (x'x)^2$	Similar to most of the unselected examples

Experimental Testbeds

1. Five datasets from the UCI machine learning repository

Experimental Testbeds

1. Five datasets from the UCI machine learning repository
2. Medical image classification, randomly select $2,785$ medical images from the ImageCLEF (Lehmann et al., 2005) that belong to 150 different categories. Each image is represented by $2,560$ visual features.

$F1$ metric

Use classification $F1$ performance as evaluation metric.

$$F1 = 2 * p * \frac{r}{p + r} \tag{18}$$

Harmonic mean of precision p and recall r of classification.

Large Margin Classifiers

Two large margin classifiers are used as the basis classifiers:

1. Kernel logistic regressions (KLR-AL) (Zhu & Hastie, 2001)
    - Measures classification uncertainty based on entropy of distribution $p(y|x)$
    - Selects examples with largest entropy for manual labeling

Large Margin Classifiers

Two large margin classifiers are used as the basis classifiers:

1. Kernel logistic regressions (KLR-AL) (Zhu & Hastie, 2001)
   - Measures classification uncertainty based on entropy of distribution $p(y|x)$
   - Selects examples with largest entropy for manual labeling
2. Support vector machine active learning (SVM-AL) (Tong & Koller, 2000)
   - Determines classification uncertainty of an example x by its distance from the decision boundary $x^T x + b = 0$
   - Selects examples with smallest distance

Evaluate Performance of Competing Active Learning Algorithms

1. Randomly pick l training samples from dataset for each category s.t. $\#$ negative examples $=$ $\#$ positive examples

Evaluate Performance of Competing Active Learning Algorithms

1. Randomly pick l training samples from dataset for each category s.t. # negative examples = # positive examples
2. Train SVM and KLR classifiers using the l labeled examples

Evaluate Performance of Competing Active Learning Algorithms

1. Randomly pick l training samples from dataset for each category s.t. # negative examples = # positive examples
2. Train SVM and KLR classifiers using the l labeled examples
3. Additional s ("batch size") unlabeled examples are chosen for manual labeling for each AL method

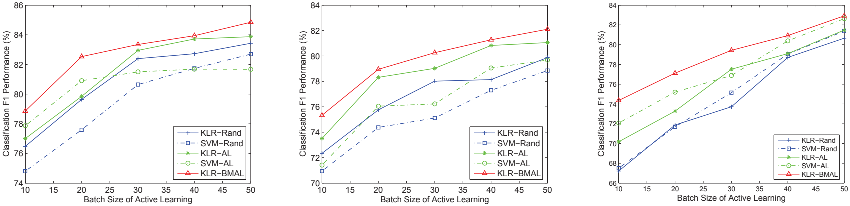
Evaluate Performance of Competing Active Learning Algorithms

1. Randomly pick l training samples from dataset for each category s.t. # negative examples = # positive examples
2. Train SVM and KLR classifiers using the l labeled examples
3. Additional s ("batch size") unlabeled examples are chosen for manual labeling for each AL method
4. For comparison, train two reference models by randomly selecting s samples for manual labeling (SVM-Rand & KLR-Rand)

Evaluate Performance of Competing Active Learning Algorithms

1. Randomly pick l training samples from dataset for each category s.t. # negative examples = # positive examples
2. Train SVM and KLR classifiers using the l labeled examples
3. Additional s ("batch size") unlabeled examples are chosen for manual labeling for each AL method
4. For comparison, train two reference models by randomly selecting s samples for manual labeling (SVM-Rand & KLR-Rand)

Classification F1 Performance on UCI datasets



(a) Australian (b) Heart (c) Sonar

Figure 2. Evaluation of classification F1 performance on the UCI datasets with different batch sizes.

Evaluation of classification F1 performance on UCI datasets

Batch Mode Active Learning and Its Application to Medical Image Classification

Table 3. Evaluation of classification F1 performance on the UCI datasets.

Dataset	Active Learning Iteration-1					Active Learning Iteration-2				
	SVM-Rand	KLR-Rand	SVM-AL	KLR-AL	KLR-BMAL	SVM-Rand	KLR-Rand	SVM-AL	KLR-AL	KLR-BMAL
Australian	74.80 ±1.97	76.48 ±2.16	77.86 ±0.84	77.00 ±1.14	**78.86** ±1.00	79.29 ±1.30	80.89 ±1.29	80.73 ±0.93	81.43 ±0.89	**83.49** ±0.36
Breast	96.34 ±0.37	96.10 ±0.33	96.80 ±0.20	97.05 ±0.02	**97.67** ±0.06	96.80 ±0.23	96.26 ±0.55	97.52 ±0.07	97.71 ±0.06	**97.81** ±0.03
Heart	70.94 ±1.29	72.34 ±1.46	71.41 ±2.39	73.51 ±1.80	**75.33** ±1.26	76.76 ±0.70	77.84 ±0.78	76.92 ±0.91	78.78 ±1.12	**79.53** ±0.59
Ionosphere	88.58 ±0.83	88.78 ±0.81	89.05 ±1.12	89.66 ±1.10	**92.39** ±0.69	90.45 ±0.59	90.60 ±0.61	93.42 ±0.51	93.71 ±0.49	**94.26** ±0.55
Sonar	67.51 ±1.57	67.22 ±1.49	72.07 ±0.84	70.18 ±1.28	**74.36** ±0.43	73.80 ±0.81	73.33 ±0.97	75.11 ±0.87	74.80 ±0.78	**77.49** ±0.45

Conclusion

- Use batch mode active learning to select multiple examples for labeling

Conclusion

- Use batch mode active learning to select multiple examples for labeling
- Use Fisher information matrix to measure model uncertainty & choose set of examples that effectively reduce the Fisher information

Conclusion

- Use batch mode active learning to select multiple examples for labeling
- Use Fisher information matrix to measure model uncertainty & choose set of examples that effectively reduce the Fisher information
- Solve related optimization problem with an efficient greedy algoirthm that approximates the objective function by a submodular function

Conclusion

- Use batch mode active learning to select multiple examples for labeling
- Use Fisher information matrix to measure model uncertainty & choose set of examples that effectively reduce the Fisher information
- Solve related optimization problem with an efficient greedy algoirthm that approximates the objective function by a submodular function
- Experical studies show method to be more effective than margin-based active learning approaches