Batch Mode Active Learning

Rong Jin
Department of Computer Science and Engineering
Michigan State University, USA

Joint work with Steven H. Hoi* and Michael R. Lyu+

*School of Computer Engineering, Nanyang Technological University
+ Department of Computer Science and Engineering, Chinese University of Hong Kong
Active Learning: Introduction

• Goal:
  • Learn an accurate classification model by a small number of labeled examples

• Key idea:
  • Label the most informative examples

• Key question:
  • How to identify the most informative examples?
  • *Uncertainty principle*: the most uncertain examples are the most informative ones
Active Learning: Related Work

• Key issue: measure classification uncertainty
  • Ensemble approach (e.g., query by committee [Freud et al., 1997])
  • Classification margin (e.g., SVM based active learning [Tong and Koller, 2000])
  • Measure the variance of classification models (e.g., experimental design [Yu et al., 2006])
  • Bayesian approaches that take into account both classification uncertainty and model uncertainty [Jin et al., 2003]
Active Learning for Image Retrieval

• Image retrieval as binary classification
  • Given a query, present user with a few *informative* images
  • Learn *a classification model* from the feedbacks
  • Use the learned model as the *refined query*

• Compared to relevance feedback
  • Most *similar* images vs. most *informative* images
  • Also referred as “*active feedback*”
Active Learning for Image Retrieval

• Limitation
  • Most active learning algorithms are designed to choose *one* example at one time
  • *Multiple* images are returned for each query for relevance judgments

• Solution: *Batch Mode Active Learning*
  • Select *multiple informative* and *diverse* examples simultaneously
Batch Model Active Learning

- Challenge: *redundant* examples

- Positive examples of class-1
- Negative examples of class-2
- Unlabeled examples
- Selected examples for labeling

---

D1: learned decision boundary  
D2: true decision boundary

(a) Binary classification example  
(b) Margin-based active learning  
(c) Batch mode active learning
Fisher Information Matrix

- Measure the model uncertainty
  - Cramer-Rao bound
- Given a data distribution $q(x)$ and a classification model $p(y|x)$, \textit{Fisher information matrix} is defined

\[
I_q(\alpha) = \mathbb{E}_{x,y} \left[ -\frac{\partial^2}{\partial \alpha^2} \log p(y|x) \right]
\]

\[
= - \int q(x) \sum_{y=\pm 1} p(y|x) \frac{\partial^2}{\partial \alpha^2} \log p(y|x) dx
\]

$\alpha$: the mode parameter.
Logistic Regression Model

- Logistic regression model

\[ p(y|x) = \frac{1}{1 + \exp(-y\alpha^\top x)} \]

\( \alpha \): the mode parameter needs to be learned.

- Fisher information matrix

\[ I_p(\alpha) = \sum_x q(x)\pi(x)(1 - \pi(x))xx^\top + \delta I \]

\[ \pi(x) = \frac{1}{1 + \exp(\alpha^\top x)} = p(y = -1|x) \]
Kernel Logistic Regression (KLR)

Given labeled examples

\[ \mathcal{L} = \{(y_1, x_1^l), (y_2, x_2^l), \ldots, (y_m, x_m^l)\} \]

and a kernel function \( K(x, x') \), according to the representer theorem, we have

\[ w^T x \rightarrow K(w, x) = \sum_{x' \in \mathcal{L}} \theta(x') K(x', x) \]

\[ p(y|x) = \frac{1}{1 + \exp \left( -y \sum_{x' \in \mathcal{L}} \theta(x') K(x', x) \right)} \]
Theoretical Foundation

Maximize Fisher information matrix [Zhang, 2003]

\[
S^* = \arg \min_{|S|=k} \text{tr}(I_q(S, \alpha)^{-1} I_p(\alpha))
\]

\[p(x) : \text{distribution of unlabeled examples}\]

\[p(x) = \frac{1}{|U|} \sum_{x_i \in U} \delta(x - x_i)\]

\[k : \text{the set of selected examples}\]

\[q(x) : \text{distribution of selected examples } S\]

\[q(x) = \frac{1}{|S|} \sum_{x_i \in S} \delta(x - x_i)\]
Why Fisher Information Matrix?

\[ S^* = \arg \min_{S \subseteq \mathcal{U} \land |S|=k} \text{tr}(I_q(S, \alpha)^{-1} I_p(\alpha)) \]

Let \( S = (x_1, \ldots, x_k) \) be selected examples

Classification uncertainty

\[
-\text{tr}(I_q(S, \alpha)^{-1} I_p(\alpha)) \propto \sum_{j=1}^{n} \pi_j (1 - \pi_j) \log \mathcal{N}(x_j | m, \Lambda)
\]

\[
m = \sum_{i=1}^{k} \gamma_i x_i, \quad \Lambda = \delta I + \sum_{i=1}^{k} \gamma_i (x_i - m)(x_i - m)^\top, \quad \gamma_i = \frac{\pi_i (1 - \pi_i)}{\sum_{j=1}^{k} \pi_j (1 - \pi_j)}
\]
Why Fisher Information Matrix?

\[ S^* = \arg \min_{S \subseteq U \land |S| = k} \text{tr}(I_q(S, \alpha)^{-1}I_p(\alpha)) \]

Let \( S = (x_1, \ldots, x_k) \) be selected examples

\[-\text{tr}(I_q(S, \alpha)^{-1}I_p(\alpha)) \propto \sum_{j=1}^{n} \pi_j (1 - \pi_j) \log N(x_j | m, \Lambda)\]

\[m = \sum_{i=1}^{k} \gamma_i x_i, \quad \Lambda = \delta I + \sum_{i=1}^{k} \gamma_i (x_i - m)(x_i - m)^\top, \quad \gamma_i = \frac{\pi_i (1 - \pi_i)}{\sum_{j=1}^{k} \pi_j (1 - \pi_j)}\]
Optimization Problem

Select the $k$ most informative examples

$$S^* = \arg \min_{S \subseteq U \land |S| = k} \text{tr}(I_q(S, \alpha)^{-1}I_p(\alpha))$$

- A combinatorial optimization problem
- Need efficient algorithms to find the approximate solution
Relax $q(x)$

$$q(x) = \sum_{x_i \in \mathcal{U}} q_i \delta(x - x_i)$$

$$q_i = \begin{cases} 
\frac{1}{k} & \text{if } x_i \in \mathcal{S} \\
0 & \text{if } x_i \notin \mathcal{S} 
\end{cases}.$$  

Relaxation

$$\sum_{x_i \in \mathcal{U}} q_i = 1$$

$$q_i \in [0, 1]$$

$$\min_{q, M} \text{tr}(M)$$

s. t.  

$$M \succeq I^{1/2}_p I^{-1}_q I^{1/2}_p$$

$$\sum_{i=1}^{n} q_i = 1, q_i \geq 0, i = 1, \ldots, n$$
Relax $q(x)$

$$\min_{q,M} \quad \text{tr}(M)$$

s. t. $$\left( \begin{array}{cc} \sum_{i=1}^{n} q_i \pi_i (1 - \pi_i) x_i x_i^T & I_p^{1/2} \\ I_p^{1/2} & M \end{array} \right) \succeq 0$$

$$\sum_{i=1}^{n} q_i = 1, \quad q_i \geq 0, \quad i = 1, \ldots, n$$

- Semi-definitive programming
  - Computationally expensive for a large number of examples

- Key idea of improving efficiency
  - Matrix inequality $\rightarrow$ a finite number of nonlinear constraints
Eigen Space Simplification

Assume \( M = \sum_{i=1}^{s} \gamma_i v_i v_i^\top \)

• \((\lambda_i, v_i), i = 1, \ldots, s\) are principle eigen-vectors and eigenvalues of \(I_p\)

\[ M \succeq I_p^{1/2} I_q^{-1} I_p^{1/2} \]

Linear Matrix Inequality

\[ v_k^\top I_q v_k \geq \gamma_k^{-1} \lambda_k, \quad k = 1, \ldots, s \]

Nonlinear constraints
Eigen Space Simplification

\[
\min_{q, M} \quad \text{tr}(M)
\]

s. t. \( M \succeq I_p^{1/2} I_q^{-1} I_p^{1/2} \)

\[
\sum_{i=1}^{n} q_i = 1, q_i \geq 0, i = 1, \ldots, n
\]

\[
\min_{q \in \mathbb{R}^n} \quad \sum_{k=1}^{s} \frac{\lambda_k}{\sum_{i=1}^{n} q_i \pi_i (1 - \pi_i) (x_i^T v_k)^2}
\]

s.t. \( \sum_{i=1}^{n} q_i = 1, q_i \geq 0, i = 1, \ldots, n \)

How to Optimize?
Bound Optimization

- Upper bound the objective function
- Iterative procedure
- It guarantees the global optimal solution

\[
q_{i}^{t+1} \leftarrow \left[ q_{i}^{t} \right]^{2} \pi_{i} (1 - \pi_{i}) \sum_{k=1}^{s} \frac{(x_{i}^{T}v_{k})^{2} \lambda_{k}}{\left( \sum_{j=1}^{n} q_{j}^{t} \pi_{j} (1 - \pi_{j})(x_{j}^{T}v_{k})^{2} \right)^{2}}
\]

Solution of the next iteration

Solution of current iteration
Testbed

• Database: 5000 Corel images from 50 categories
• Each image is represented by 36 features
  • Color, edges, and textures
• Randomly choose 100 images as queries
• Relevance judgments based on image categories
  • An returned image is relevant if it belongs to the same category as the query image
• Evaluation metric: average precision
Experimental Results

Present users with randomly selected images
Experimental Results

Present users with images with the largest uncertainty.
Experimental Results

Present users with images selected by Batch Model Active Learning

Present users *sequentially* images with the largest uncertainty
Conclusion

• Present the study of batch mode active learning
• Key challenge: find informative and diverse examples
• A framework of batch mode active learning based on the Fisher information matrix
• Efficient algorithms based on bound optimization
• Encouraging results on image retrieval
• Future work
  • Cost sensitive batch mode active learning