Active Learning with Hinted Support Vector Machine

Chun-Liang Li  Chun-Sung Ferng  Hsuan-Tien Lin

National Taiwan University

2012/11/6
Given

- The labeled pool $D_l = \{(feature \ x_i, label \ y_i)\}_{i=1}^{N}$, $y_i \in \{+1, -1\}$
- The unlabeled pool $D_u = \{\tilde{x}_j\}_{j=1}^{M}$

A pool-based active learning algorithm iteratively

- use querying algorithm $Q$ to query $\tilde{x}_s \in D_u$
- update $D_l$ and $D_u$
- learn a decision function $f^{(r)}$ by learning algorithm $L$

and improve the performance of $f^{(r)}$ w.r.t #queries

Goal

use few queries to improve performance of decision function
Uncertainty Sampling (A Popular Paradigm)

In each iteration, query the least certain one

Tong and Koller (2000)

- learn a SVM hyperplane for choosing the instance closest to the boundary
- use the same hyperplane for querying and learning

- blue framed: labeled instances
- magenta circled: to be queried
Potential Drawback

(a) Initial Stage

(b) After #iterations

be overly confident to unknown area

Representative Sampling

- clustering-based algorithms (Donmez et al., 2007)
- label estimation in semi-supervised learning (Huang et al., 2010)
Hinted Sampling

Intuition

Use some unlabeled instances $D_h \subseteq D_u$ as hints (Abu-Mostafa, 1995) to make querying boundary be aware of (pass through) unknown areas.

(c) Initial Stage

(d) After #iterations

querying boundary is different from the decision boundary (black)
Active Learning with Hinted SVM (ALHS)

- Separate $\mathcal{D}_l \rightarrow$ classification problem
- Pass through $\mathcal{D}_h \rightarrow$ regression problem

HintSVM (For querying)

$$\min_{w, b, \xi, \tilde{\xi}, \tilde{\xi}^*} \frac{1}{2} w^T w + C_l \sum_{i=1}^{\mid \mathcal{D}_l \mid} \xi_i + C_h \sum_{j=1}^{\mid \mathcal{D}_h \mid} (\tilde{\xi}_j + \tilde{\xi}_j^*)$$

subject to

- $y_i (w^T x_i + b) \geq 1 - \xi_i$ for $(x_i, y_i) \in \mathcal{D}_l$,
- $w^T \tilde{x}_j + b \leq \epsilon + \tilde{\xi}_j$ for $\tilde{x}_j \in \mathcal{D}_h$,
- $-(w^T \tilde{x}_j + b) \leq \epsilon + \tilde{\xi}_j^*$ for $\tilde{x}_j \in \mathcal{D}_h$.

- A convex optimization problem
- Uncertainty sampling with SVM is a special case of ALHS ($C_h = 0$)
Our algorithm ALHS iteratively

- select $\mathcal{D}_h \subseteq \mathcal{D}_u$
- use HintSVM in querying algorithm $Q$ to query $\tilde{x}_s \in \mathcal{D}_u$
- update $\mathcal{D}_l$ and $\mathcal{D}_u$
- learn a typical SVM $f(r)$ as decision function by learning algorithm $\mathcal{L}$
Comparison and Contribution

### Comparison

<table>
<thead>
<tr>
<th>Uncertainty Sampling</th>
<th>Querying Algo $\mathcal{Q}$</th>
<th>Learning Algo $\mathcal{L}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncertainty Sampling</td>
<td>Typical SVM</td>
<td>Typical SVM</td>
</tr>
<tr>
<td>ALHS</td>
<td>HintSVM</td>
<td>Typical SVM</td>
</tr>
</tbody>
</table>

### Contributions

- Resolve potential drawback of uncertainty sampling
- Convex Optimization - Simpler than some representative algos
- Achieve better performance
Hint Selection Strategies in ALHS

- HintSVM: \[
\min \frac{1}{2} w^T w + C_l \sum_{i=1}^{\left|D_l\right|} \xi_i + C_h \sum_{j=1}^{\left|D_h\right|} \left(\tilde{\xi}_j + \tilde{\xi}_j^*\right)
\]

- Balance cost parameters \( C_l = \max \left(\frac{\left|D_h\right|}{\left|D_l\right|}, 1\right) \times C_h \)

Hint Dropping

Too many hints would overwhelm HintSVM

- hints surrounding to a labeled instance are less useful
- Drop all \( D_h \) after \( T \) iterations (similar to Donmez et al., 2007)
Experiment I

The datasets that **representative sampling** outperforms **uncertainty sampling**

ALHS can compete with or even outperforms them
The datasets that uncertainty sampling outperforms representative sampling

ALHS can compete with or even outperforms them
Conclusion

- **general framework** of active learning: Hinted Sampling
- HintSVM: convex optimization, **simpler**
- good experimental results
- future work: a **new direction** for theoretical analysis of representative sampling

Thanks, any question?