Active Learning Strategies Using SVMs

Ming-Hen Tsai
Department of Computer Science
National Taiwan University

Joint work with Chia-Hua Ho and Chih-Jen Lin
Special Session: Active and Autonomous Learning,
IJCNN’10, July, 22th, 2010
Outline

- Introduction
- Classification Tools
- Learning Methods
- Experimental Results
- Conclusions
Outline

- Introduction
- Classification Tools
- Learning Methods
- Experimental Results
- Conclusions
The Problem

- Only one instance is labeled at first.
- We can query any number of labels in training data.
- **Goal:** Achieve high AUC (Area Under Curve) with as few queried samples as possible.
- Evaluated by ALC (Area under Learning Curve).
Our Framework

Obtain a classifier (One labeled point)

Predict and Query

Obtain a classifier

- After we have more than one labeled instance, only labeled instances are used.
- Query method is the same in the whole procedure.
Outline

- Introduction
- Classification Tools
- Learning Methods
- Experimental Results
- Conclusions
Classification Tools

- Support Vector Classification
- Support Vector Regression
- Logistic Regression
- Transductive Support Vector Machine
- One-class Support Vector Machine
Support Vector Classification

- Proposed by Bernhard E. Boser, Isabelle Guyon and Vladimir Vapnik.
- Training instances: \( \{y_i, x_i\} \), \( y_i = \pm 1 \), \( i = 1, 2, \ldots, l \).
- SVC solves

\[
\min_{w,b} \quad \frac{1}{2} \|w\|^2_2 + C \sum_{i=1}^{l} \max \left( 1 - y_i (w^T \phi(x_i) + b), 0 \right)
\]

- \( w \) is the weight vector.
Logistic Regression

- Training instances: \( \{y_i, \mathbf{x}_i\}, y_i = \pm 1, \ i = 1, 2, \ldots, l \).
- L1-regularized LR (L1 LR) solves
  \[
  \min_{\mathbf{w}} \|\mathbf{w}\|_1 + C \sum_{i=1}^{l} \xi(\mathbf{w}, y_i, \mathbf{x}_i)
  \]
- L2-regularized LR (L2 LR) solves
  \[
  \min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^{l} \xi(\mathbf{w}, y_i, \mathbf{x}_i)
  \]
- Logistic loss function:
  \[
  \xi(\mathbf{w}, y_i, \mathbf{x}_i) = \log (1 + \exp (-y_i \mathbf{w}^T \mathbf{x}_i)).
  \]
Transductive SVM (1/2)

- Semisupervised classifier
- Labeled training instances: \( \{ y_i, x_i \} \), \( y_i = \pm 1 \), \( i = 1, 2, \ldots, l \); unlabeled features values: \( x'_i \in \mathbb{R}^n \), \( i = 1, 2, \ldots, u \)

TSVM solves

\[
\min_{w, \{y'_i\}_1^u} \frac{\lambda}{2} \|w\|_2^2 + \frac{1}{2l} \sum_{i=1}^{l} \xi^2(w, y_i, x_i) + \frac{\lambda'}{2u} \sum_{i=1}^{u} \xi^2(w, y'_i, x'_i)
\]
L2-loss functions and constraints:

\[ \xi(w, y, x) = \max \left( 1 - y(w^T x), 0 \right), \]
\[ \frac{1}{u} \sum_{i=1}^{u} \max \left( 0, \text{sgn}(w^T x'_j) \right) = r, \]
\[ y'_j \in \{-1, 1\} \]

- \( r \) is the ratio of positive unlabeled instances.
Software

- LIBSVM [Chang and Lin, 2001]: SVC (RBF kernel), SVR
- LIBLINEAR [Fan et al., 2008]: SVC (linear kernel), LR
- SVMlin [Sindhwani and Keerthi, 2006]: TSVM
Outline

- Introduction
- Classification Tools
- Learning Methods
- Experimental Results
- Conclusions

(National Taiwan Univ.)

July 22, 2010
Training with One Labeled Instance

Obtain a classifier (One labeled point)

Predict and Query

Obtain a classifier

- Only applied when there is only one labeled training instance.
Transductive SVM accepts unlabeled instances.

Parameter $r$ (positive ratio in unlabeled data) is unknown.

Use the positive ratio of similar development data sets or a small constant ratio.
The positive ratios in all development data sets are small.

Assign all unlabeled instances negative labels and a small weight.

→ There are one positive instance and many negative instances.
Querying Strategies

- Obtain a classifier (One labeled point)
- Predict and Query
- Obtain a classifier

- Query points which improve the AUC.
- We only use the predicted decision values to decide which points to query.
Arcsin-discretized Query (1/2)

- Query many instances close to the decision boundary and also query some instances far from the decision boundary.
- Scale decision values to $[-\pi/2, \pi/2]$.
- Query $n + 1$ instances whose decision values are close to

$$\sin^{-1}\left(-1 + \frac{0}{n}\right), \sin^{-1}\left(-1 + \frac{2}{n}\right),$$

$$\sin^{-1}\left(-1 + \frac{4}{n}\right), \ldots, \sin^{-1}\left(-1 + \frac{2n}{n}\right).$$
Arcsin-discretized Query (2/2)
Passive Learning (1/2)

- Query all training instances in the first time.
- Used when active learning methods do not seem to improve AUC significantly.
- It is a good method if the learning curve is convex (e.g., ORANGE).
- It should not be applied if the learning curve is concave (e.g., NOVA).
Passive Learning (2/2)

[Graph showing area under the ROC curve vs. log number of labels queried for different learning methods and datasets]
Other Methods

- **Simple Decision Value Method**
  - [Tong and Koller, 2002]
  - Query the method with small absolute decision values.

- **Uniform-discretized Query**
  - Query instances uniformly with respect to decision values.

- **Random Query**
  - Query instances by random.
Obtain a Classifier

- Use LR for categorical data sets.
- Use SVC with RBF kernel for non-categorical data sets.
For passive learning, we query all the instances.

Otherwise, double the number of samples every time, e.g., 16, 32, 64, ... , until all training instances are labeled.
Overview

- Introduction
- Classification Tools
- Learning Methods
- Experimental Results
- Conclusions
The development data sets and the challenge data sets are in similar domains.

Apply the best method of a development data set on the similar challenge data set.
Construct a 1-1 mapping on development and challenge data sets:

- A → IBN_SINA
- B → ORANGE
- C → HIVA
- D → NOVA
- E → ZEBRA
- F → SYLVA
Preprocessing

- Missing values are filled by 0.
- Scaled to $[0, 1]$ except IBN_SINA and A.
- For ORANGE and B, missing value indicators are added.
  
  e.g., 245 in the total 250 features of B have missing values.

  Finally there are 495 features in B.

- For C, categorical features are expanded to binary features.
## Experimental Results

### AUC: The First Point

<table>
<thead>
<tr>
<th>Method</th>
<th>HIVA</th>
<th>NOVA</th>
<th>IBN</th>
<th>ORANGE</th>
<th>SYLVA</th>
<th>ZEBRA</th>
</tr>
</thead>
<tbody>
<tr>
<td>All negative</td>
<td>0.530</td>
<td>0.656</td>
<td>0.424</td>
<td>0.514</td>
<td>0.774</td>
<td>0.402</td>
</tr>
<tr>
<td>TSVM 0.1</td>
<td>0.413</td>
<td>0.356</td>
<td>0.883</td>
<td>0.563</td>
<td>0.943</td>
<td>0.707</td>
</tr>
<tr>
<td>TSVM real</td>
<td>0.432</td>
<td>0.332</td>
<td>0.823</td>
<td>0.577</td>
<td>0.932</td>
<td>0.704</td>
</tr>
</tbody>
</table>
## ALC: Query Method (categorical)

<table>
<thead>
<tr>
<th>Method</th>
<th>HIVA</th>
<th>NOVA</th>
<th>ORANGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>passive</td>
<td>0.320</td>
<td>0.643</td>
<td>0.378</td>
</tr>
<tr>
<td>random</td>
<td>0.177</td>
<td>0.677</td>
<td>0.265</td>
</tr>
<tr>
<td>simple</td>
<td>0.083</td>
<td>0.694</td>
<td>0.265</td>
</tr>
<tr>
<td>uniform</td>
<td>0.168</td>
<td>0.751</td>
<td>0.249</td>
</tr>
<tr>
<td>arcsin</td>
<td>0.133</td>
<td>0.753</td>
<td>0.226</td>
</tr>
</tbody>
</table>
### ALC: Query Method (non-categorical)

<table>
<thead>
<tr>
<th>Method</th>
<th>IBN_SINA</th>
<th>SYLVA</th>
<th>ZEBRA</th>
</tr>
</thead>
<tbody>
<tr>
<td>passive</td>
<td>0.874</td>
<td>0.941</td>
<td>0.550</td>
</tr>
<tr>
<td>random</td>
<td>0.897</td>
<td>0.940</td>
<td>0.395</td>
</tr>
<tr>
<td>simple</td>
<td>0.723</td>
<td>0.800</td>
<td>0.274</td>
</tr>
<tr>
<td>uniform</td>
<td>0.860</td>
<td>0.935</td>
<td>0.387</td>
</tr>
<tr>
<td>arcsin</td>
<td>0.900</td>
<td>0.967</td>
<td>0.308</td>
</tr>
</tbody>
</table>
Obtain a classifier (One labeled point)

Predict and Query

Obtain a classifier

A  TSVM  $r = 0.1$
B  TSVM  $r = 0.1$
C  All negative
D  All negative
E  TSVM  $r = 0.1$
F  TSVM  $r = 0.1$
Obtain a classifier (One labeled point)

Predict and Query

Obtain a classifier

A  arcsin-discretized
B  passive learning
C  passive learning
D  arcsin-discretized
E  passive learning
F  arcsin-discretized
Obtain a classifier (One labeled point) → Predict and Query → Obtain a classifier

A  SVC RBF
B  L1 LR
C  L2 LR
D  L2 LR
E  SVC RBF
F  SVC RBF
### Results of Challenge Data Set

<table>
<thead>
<tr>
<th></th>
<th>ALC</th>
<th></th>
<th>ALC</th>
<th>rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBN</td>
<td>0.900</td>
<td>SINA</td>
<td>0.550</td>
<td>3</td>
</tr>
<tr>
<td>ORANGE</td>
<td>0.378</td>
<td>B</td>
<td>0.376</td>
<td>1</td>
</tr>
<tr>
<td>HIVA</td>
<td>0.320</td>
<td></td>
<td>0.341</td>
<td>4</td>
</tr>
<tr>
<td>NOVA</td>
<td>0.753</td>
<td></td>
<td>0.665</td>
<td>3</td>
</tr>
<tr>
<td>ZEBRA</td>
<td>0.422</td>
<td></td>
<td>0.585</td>
<td>2</td>
</tr>
<tr>
<td>SYLVA</td>
<td>0.967</td>
<td></td>
<td>0.709</td>
<td>6</td>
</tr>
</tbody>
</table>
Outline

- Introduction
- Classification Tools
- Learning Methods
- Experimental Results
- Conclusions
Conclusions

- We can quickly decide the query instances after training because we only use decision values.
- We do not know how to automatically decide the number of samples to query.
- We do not know how when to conduct passive learning.
- arcsin-discretized query have good performance.
Thank you!