User Gender Prediction Using Browser Interaction Pattern

Motivation & Problem

Motivation
- User demographic profile (gender/age): important for customized web service / targeted advertising
- Hard to acquire labels directly through form-based survey: burden on users, security & privacy issues

Previous Work
- Use visited webpage/URL history, text features

Research Problems
- How to predict a user’s gender using browser UI interaction history?
- How to use the unlabeled data to improve the prediction accuracy?

Dataset

Firefox 4 Beta Interface Test Pilot dataset
- The history of browser UI interaction collected from 1,134 users (567 male, 567 female) within a test duration of 7 days
- Each sample $X$ is in the format of $(U, C, t)$
  - "User $U$ used the UI component $C$ at time $t"
- Target gender label $Y = \{\text{Male, Female}\}$

Feature Extraction & Gender Classification

UI Usage Histogram Feature
- The percentage of time spent on each of the 14 UI categories
- The percentage of time spent on each of the 94 UI items

UI Item Transition Rate Feature
- Mean, median, and standard deviation of the time difference of successive UI item click

Within-Category UI Usage Interval Feature
- Mean, median, and standard deviation of the time interval within the same UI category identifier

Dimensionality Reduction
- Use Principle Component Analysis (PCA): Reduce 139 original feature dimensions to 20 dimensions

Supervised Learning
- Compare Naive Bayes, Decision Tree, k-NN, and SVM
- Experiment setting: 10-fold cross validation

Semi-Supervised Learning
- Self-Training Algorithm
  - For each iteration train a classifier from labeled data $L$
  - Classify samples in unlabeled data $U$
  - Add $m$ most-confident classifications to $L$ ($m = 10$)
- Experiment setting: Initial $|L| = 85$, $|U| = 766$, $|\text{TestSet}| = 283$ (25%)

Experimental Results

Feature/Classifier Combinations
- $k$-NN and Decision Tree achieve 72% accuracy using all features
- UIHistogram & UIInterval more useful than UITransition

Parameter Selection
- Decision Tree Classifier
  - Parameter Selection: $\frac{1}{L}$
- k-NN Classifier
  - Parameter Selection: $k$

Semi-Supervised Learning Result
- 67% accuracy with initial 85 labeled data
- 766 unlabeled data were provided
- Converged to 72% after 50 iterations