National Institute of Advanced Industrial Science and Technology
AIST

KDDI R&D LABS

A Fast and Simple Method for Profiling a Population of Twitter Users

MUSE2012@Bristol, UK

Hideki ASOH, AIST
Kazushi IKEDA, KDDI R&D Labs, Inc.
Chihiro ONO, KDDI R&D Labs, Inc.

Background

- Twitter as a real-time marketing tool
- Most users do not disclose profile (age, sex, location, occupation, etc.)
- Estimating user profile is important for marketing applications
  - targeting advertisement

Profiling each User

- Estimating user profile from each user's tweets [Ikeda+ 2010]
  - Using SVM as classifier
  - Input: Occurrences of characteristic words in last 200 tweets of the target user
  - Training classifiers using tweets from users who disclose profile
  - Choose characteristic words for each segment using information criteria (AIC) (about 10,000users/segment)

Examples of Characteristic Words

Male	Female
Government	Husband
Android	Mother
Wife	Bath
Company	Laundry
Google	Lunch

10s	20s	30s	40s
Mathematics	University	Work	Son
School	Part-time	Company	Holiday
Examination	Seminar	Business	Golf
Test	Job-hunting	Boss	Diplomacy
Physical Ed.	Lecture	Beer	Backache

Profiling a Population of Users

- In some applications, estimating profile of each user is NOT necessary
- Profiling a population of users who tweet about a specific event, content is required
  - ratio of male/female
  - ratio of 10s/20s/30s/40s/...
- ⇒ Class ratio estimation problem

Class Ratio Estimation

- Observation: X Class: Y
- Class ratio of the training sample and test sample (= target population) are different
  - Biased training sample
- Estimating class ratio in the target population

Class Ratio Estimation

- A kind of transfer learning

  $P_{\text{train}}(X, Y) \mathrel{!}= P_{\text{test}}(X, Y)$

- Class Ratio Estimation

  $P_{\text{train}}(X|Y) = P_{\text{test}}(X|Y)$

  $P_{\text{train}}(Y) \mathrel{!}= P_{\text{test}}(Y)$

- cf. Covariate Shift [Shimodaira 2000]

  $P_{\text{train}}(Y|X) = P_{\text{test}}(Y|X)$

  $P_{\text{train}}(X) \mathrel{!}= P_{\text{test}}(X)$

Approaches

- Baseline : Classifying each sample and calculate class ratios by aggregation
  - Time consuming when the size of the target population is large
  - Classifier is suffered from sampling bias
    - $P_{train}(Y|X)$!= $P_{test}(Y|X)$
    - Prior estimation using EM algorithm and bias correction [Saerens 2002] -> more time consuming
- Direct Method : Class ratio estimation without classification of each sample

Idea of Direct Method

- Training data $P(x,y) = P(x|y)P(y)$
- Test data $Q(x,y) = P(x|y)Q(y)$

$$Q(x) \equiv \sum_c P(x|y=c)Q(y=c)$$

$$Q'(x;\theta) \equiv \sum_c P(x|y=c)\theta_c$$

- Choose optimal θc which makes $Q'(x)$ similar to $Q(x)$

Du Plessis and Sugiyama [ICML 2012]

- Minimizing *f*-divergence

$$D_f(Q(x)||Q'(x)) = \int Q(x) f\left(\frac{Q'(x)}{Q(x)}\right) dx$$

- *f*-divergence
  - KL divergence（$f(z) = \log(z)$）
  - PE divergence（$f(z) = (z-1)^2$）
- using density ratio approximation for approximating divergence

Difficulties

- Computational Cost is high
- Good for low dimensional continuous inputs
- Not good for high dimensional characteristic word vector
  - some thousands dimensional
  - discretized data

Minimizing Difference of Means

- Instead of minimizing divergence, minimizing the difference of means of Q'(x) and Q(x)

$$\overline{Q(x)} = \int xQ(x)dx$$

$$\overline{Q'(x)} = \sum_c \theta_c \int xP(x|y=c)dx$$

$$\hat{\theta}_c = \text{argmin}\left(\int xQ(x)dx - \sum_c \theta_c \int xP(x|y=c)dx \right)^2$$

Implementation

- Mean of the test dataset: μ
- Class mean of the training dataset: μ_c
- Class mean matrix: $A = (\mu_1, \mu_2, \ldots, \mu_K)$
- Class ratio vector: $\theta=(\theta_1,\ldots,\theta_K)$

$$\underset{\theta}{\operatorname{argmin}}(\mu - A\,\theta^T)^2$$

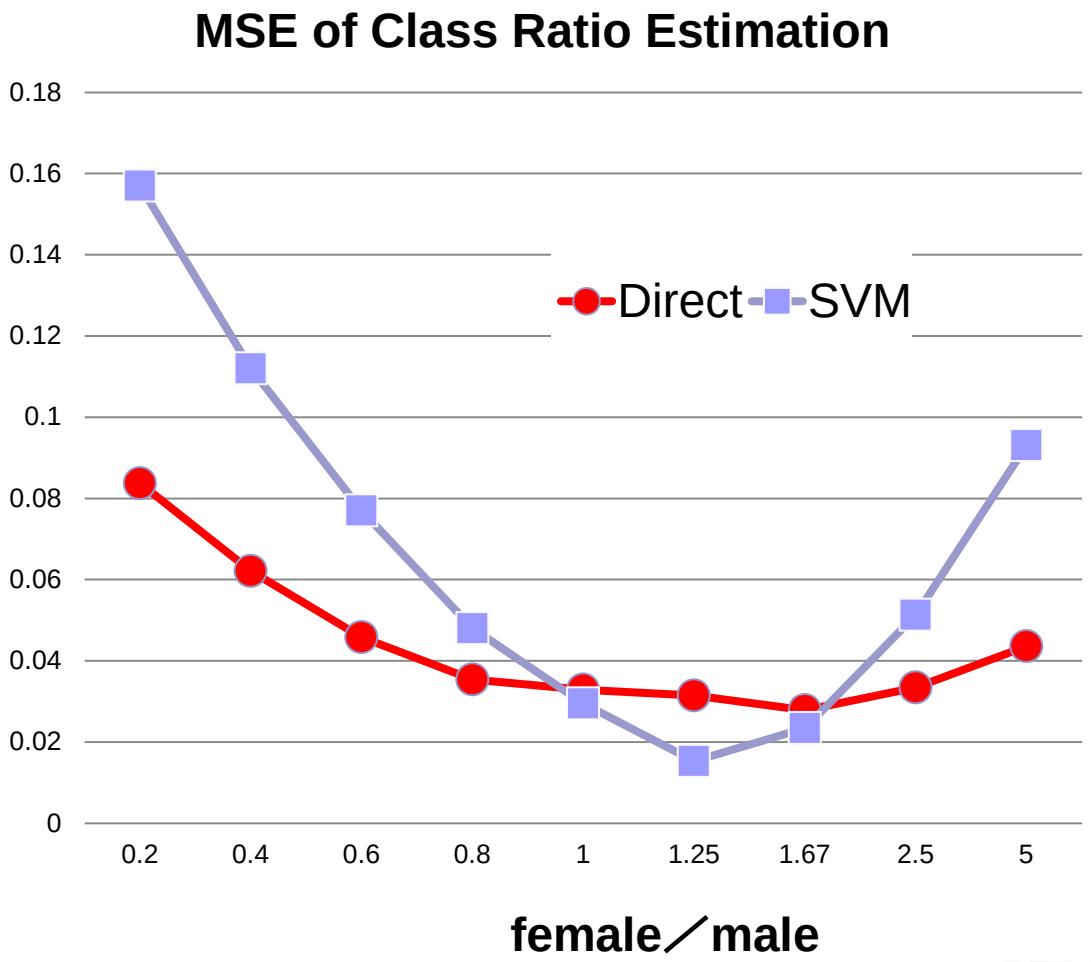
$$\theta^* = A^+\,\mu$$

Experiments

- Male/female ratio
  - 2 classes
  - 1000 users, 200 tweets per a user
  - characteristic words：4000 words
- Age class ratio
  - 4 classes（10s, 20s, 30s, over40）
  - 762 users, 200 tweets per a user
  - characteristic words: 8000 words

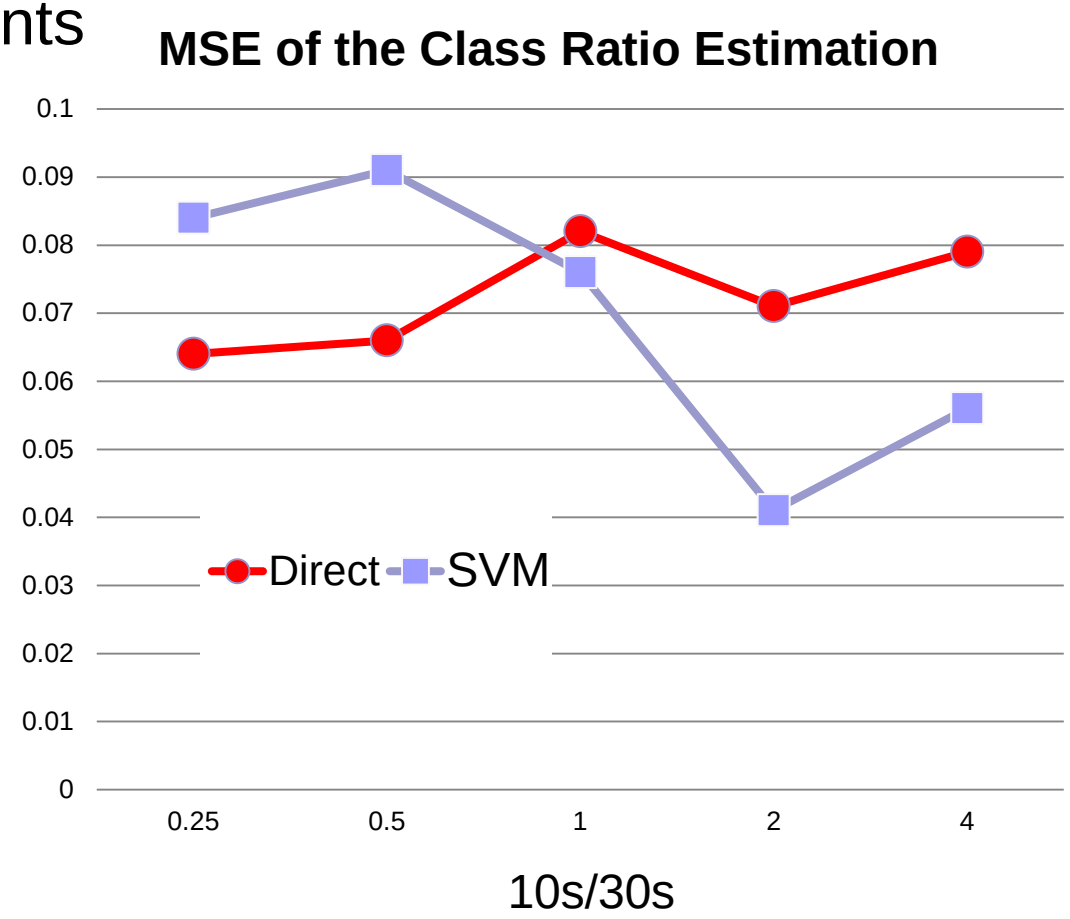
Male/Female Ratio

- Condition of Experiment
  - 600 training
    400 test
  - Training data
    female/male = 1:1
  - Control female/male
    in test data
  - Average of 10 times
- Results
  - Direct method is more stable than baseline
  - 60 times faster than baseline (in R)



Age Class Ratio

- Condition of Experiments
  - 458 training
    304 test
  - Age ratio of
    training data 1:1:1:1
  - Control 10s/30s in
    test data
  - Average of 5 times
- Results
  - Direct method is
    more stable but ...
  - Difficult to interpret



National Institute of Advanced Industrial Science and Technology
AIST

KDDI R&D LABS

Summary

- Introduce the class ratio estimation problem
- Propose a very simple direct method without estimating profile of each user
- Apply the method to Twitter data in the marketing context
- The simple direct method is competitive and can be more robust than the baseline
- Performance seems to be degraded when the number of classes increases

Future Work

- More detailed evaluations
  - Various conditions, more users
- Bias-Variance trade-off
  - Try baseline method with weaker classifiers
    - naive Bayes, logistic classifier, etc.
- Improvement of the method
  - Minimize KL divergence
  - Introduce higher order statistics
- Performance evaluation in real services
  - Evaluate speed up effect
  - Validation of the estimated results