Estimating Age on Twitter Using Self-Training Semi-Supervised SVM

Tatsuyuki Iju
Graduate School of Information Engineering, University of the Ryukyus
1 Senbaru, Nishihara-cho, Nakagami-gun, Okinawa, Japan

Satoshi Endo, Koji Yamada Naruaki Toma, Yuhei Akamine
School of Information Engineering, University of the Ryukyus
1 Senbaru, Nishihara-cho, Nakagami-gun, Okinawa, Jaapn
E-mail: k148582@ie.u-ryukyu.ac.jp Endo@ie.u-ryukyu.ac.jp
http://www.u-ryukyu.ac.jp/en/

Abstract

The estimation methods for Twitter user's attributes typically require a vast amount of labeled data. Therefore, an efficient way is to tag the unlabeled data and add it to the set. We applied the self-training SVM as a semi-supervised method for age estimation and introduced Plat scaling as the unlabeled data selection criterion in the self-training process. We show how the performance of the self-training SVM varies when the amount of training data and the selection criterion values are changed.

Keywords: Twitter, Age, Semi-supervised learning, Self-training, SVM, Plat scaling

1. Introduction

Nowadays, the use of Twitter as a social activity sensor has become popular trendy. Although it is more efficient to consider attribute differences such as user age and gender for analysis, users rarely share their personal information to the public. Therefore, a variety of methods for estimating Twitter user's attributes has been studied[1][2][3]. However, these methods require a vast amount of labeled data. Since collecting labeled data is typically a high cost work, the estimation method is efficient when unlabeled data is labeled and used as additional data. We investigate a method for building classifier by using self-training SVM, which is a combination of semi-supervised method, self-training and SVM. In this study, we formulate age estimation as a binary classification problem where user is labeled as under 30 or over 30. The method for user vectorization is simple bag-of-words model and all the tweets treated as data were Japanese tweets. In section 2, we describe self-training SVMs. In section3, we describe the experiments and present and analyze the results. Finally, we describe our findings and the extension of this work.

2. Self-training SVM algorithm

We describe about the method of building our classifier by using a self-training SVM. Self-training is a simple semi-supervised learning algorithm, with examples of applications started by Scudder[4]. The standard approach for self-training is as follows.

i. By Using underlying learning algorithm, train a classifier from the labeled data set.
ii. Label a part of the unlabeled data set according to the classifier, and retrain it, with the newly labeled data as an additional training set.

We construct our classifier for Twitter user's age estimation by using a self-training SVM, in which the underlying learning algorithm is SVM. Furthermore, we introduce Plat scaling as a criterion for selecting users appropriately from the unlabeled set in order to be

labeled. It is expected that poor quality data is filtered out by that criterion. Plat scaling[5] is a method used for modeling a function that returns posterior probability $P(Age|X)$ in which user X is in the Age class, according to classifier's the decision function.

The steps of the self-training SVM algorithm where sample set $\{X_i, i = 1, \dots, N_I\}$ and sample set $\{X_u, u = 1, \dots, N_U\}$ belongs to training set F_I and unlabeled set F_U, respectively.

i. Using F_I, we train a SVM and obtain probability $P(Age_l|X_u) \in \{P(Age_l|X_u), l = 1, \dots, N_A\}$ that sample X_u belongs to Age_l by Plat scaling.
ii. Define F_{u^*} containing all $X_u$with at least one of the $P(Age_l|X_u) \geq threshold$. Furthermore, Define sample in F_{U^*} as X_{u^*}.
iii. Define newF_I as $F_I = F_I + F_{U^*}$. The label of X_{u^*} is predicted as the Age_l for which $P(Age_l|X_{u^*})$ is highest. Furthermore, Define new F_U as $F_U = F_U - F_{U^*}$.
iv. Repeat i. ~ iii. until F_{U^*} cannot be defined.

performance with the test set. Additionally, in order to provide the baselines for these classifiers, we built other classifiers by using normal SVM for each of four training set arrays and measured their performance the same way, we did for the self-training SVM ones. The performance measurement was done by the five-fold cross validation. The number of users in the test set was 480 and the number of users in the unlabeled set was 1200, In addition, the number of users from under 30 and over 30 classes were balanced in both the training and test sets. The way for user vectorization was simple bag-of-words model, at the time of classifier training and user's age prediction. Prior to proceeding the experiments, it was necessary to set hyper parameters of SVM and features in a bag-of-words representation. Therefore, we performed a grid-search, set the kernel and cost parameters of SVM as linear and 1000 respectively. As for the features, we set the 158 top-ranked words by χ^2 score which appeared in user tweets from the training set.

3.2. Results

Table 1: Results for the normal SVM

	Number of the users in training set					
	76		256		376	
Age class	Under 30	Over 30	Under 30	Over 30	Under 30	Over 30
Precision	0.613	0.792	0.638	0.825	0.647	0.834
Recall	0.882	0.441	0.896	0.492	0.898	0.509
F measure	0.723	0.563	0.745	0.616	0.752	0.632
Mean F measure	0.643		0.680		0.692	

3. Experiments

3.1. Experimental contents and settings

We defined age as two classes, under 30 and over 30. This definition is the same as Rao's research[1]. we carried out some experiments in order to evaluate the performance of the Self-training SVM for the Twitter user's age estimation. By using self-training SVM, we built a classifier for each of the nine training set arrays (three sets of different number of users containing 76, 256 and 376 with three possible selection criterions of 0.5, 0.7 and 0.9 respectively). Then, we measure the classifier's

Table2: Results for the training set with 76 number of users

	Unlabeled data selection criterion					
	0.5		0.7		0.9	
Age class	Under 30	Over 30	Under 30	Over 30	Under 30	Over 30
Precision	0.614	0.792	0.610	0.793	0.625	0.791
Recall	0.882	0.442	0.888	0.430	0.901	0.473
F measure	0.723	0.563	0.722	0.554	0.728	0.590
Mean F measure	0.643		0.639		0.659	

Table3: Results for the training set with 256 number of users

	Unlabeled data selection criterion					
	0.5		0.7		0.9	
Age class	Under 30	Over 30	Under 30	Over 30	Under 30	Over 30
Precision	0.638	0.825	0.630	0.818	0.645	0.808
Recall	0.896	0.492	0.895	0.473	0.876	0.512
F measure	0.745	0.616	0.739	0.599	0.742	0.629
Mean F measure	0.680		0.669		0.686	

Table4: Results for the training set with 376 number of users

	Unlabeled data selection criterion					
	0.5		0.7		0.9	
Age class	Under 30	Over 30	Under 30	Over 30	Under 30	Over 30
Precision	0.647	0.834	0.649	0.825	0.657	0.829
Recall	0.898	0.510	0.891	0.516	0.890	0.534
F measure	0.752	0.632	0.750	0.634	0.756	0.650
Mean F measure	0.692		0.692		0.703	

First, we describe about the results for the normal SVMs. Table 1 shows the performance of the normal SVM classifiers. From Table 1, we can assure that the classifiers can improve their performance as the size of the set is increased, at least when the size of the given training set is in the 76 to 376 range. According to this result, as well as for self-training SVM, it is expected that performance can be improved by utilizing unlabeled set with the training set size within the same range size. Additionally, regardless of the training set size, it was observed that precision is high and recall is low for the under 30 class and, oppositely the precision is low and recall is high for the over 30 class. F measure was better for the under 30 class than for over 30 one. It is, basically, that the under 30 class user was easier to predict than the over 30 class one. As for the results about self-training SVM, Table 2, 3, 4, and correspond to the prediction results from classifiers built from 76, 256 and 376 users training sets, respectively. The classifier with the highest improvement from baseline was the one with 76 users training set with a selection criterion of 0.9 (improvement was 0.032 point for the over30 class recall). Classifier from the 376 users with a selection criterion of 0.9 was the second highest (improvement was 0.025 point for the over 30 class recall). Although the recall improvement for over 30 class is remarkable, precision for the over 30 class and both precision and recall for the under 30 class subtly improve or diminish as the mean F measure

Table 5: Results for the training set 76 number of users (labeling)

	Unlabeled data selection criterion					
	0.5		0.7		0.9	
Age class	Under 30	Over 30	Under 30	Over 30	Under 30	Over 30
Precision	0.621		0.722	0.699	0.752	0.705
Recall	1.0	0	0.890	0.425	0.90	0.440
F measure	0.766		0.796	0.524	0.820	0.537
Mean F measure			0.660		0.678	
Labeling rate	1.0		0.97		0.83	

merely improves. Table 5 shows the labeling accuracy for the 76 users set self-training and ratio of the labeled set for all the unlabeled set. Precision, F and mean F measures for the over 30 class were not defined, because the entire unlabeled set is filtered and eventually labeled as the under 30 class. We could observe better labeling accuracy for higher selection criterions and, in contrast, for bigger selection criterions the amount of labeled data from the unlabeled set is reduced. However, as indicated by Table 2 through Table 4, performance was better for a selection criterion of 0.5 than 0.7 for all of classifiers excluding the 376 users training set, although performance for a selection criterion of 0.9 was best for all classifiers. This result is inconsistent with the fact that the labeling error is more frequent for selection criterion of 0.5 than 0.7 as indicated in Table 5. For that reason, it is implicated that users are easier to predict if selected from a self-training process with a selection criterion of 0.7 than a 0.5 one, since the classifier with a selection criterion of 0.7 is more strongly affected by labeling error than a 0.5 one. In addition, it is inferable that, for a selection criterion of 0.9, the labeling error rate is so small that the classifier achieves in improving its performance.

4. Conclusions

In order to evaluate self-training SVM for Twitter user's age estimation, we construct a classifier by for each of the twelve training set arrays of (three sets of different users containing 76, 256 and 376 with three possible selection criterion of 0.5, 0.7 and 0.9 respectively). Then we evaluate the performance of the classifiers with test set. As a result, in the recall for the over 30 class, it was observed a 0.032 and 0.025 point of improvement from baseline for training set with a size of 76 and 376 respectively with a selection criterion of 0.9. For future works, we will investigate the relation between the selection criterion and performance of self-training SVM as well as explore the way to improve them.

References

1. Rao, Delip, et al. "Classifying latent user attributes in twitter." Proceedings of the 2nd international workshop on Search and mining user-generated contents. ACM, 2010.
2. Burger, John D., et al. "Discriminating gender on Twitter." *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011.
3. Pennacchiotti, Marco, and Ana-Maria Popescu. "Democrats, republicans and starbucks afficionados: user classification in twitter." *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011.
4. Scudder, Henry J. "Probability of error of some adaptive pattern-recognition machines." *Information Theory, IEEE Transactions on* 11.3 (1965): 363-371.
5. Platt, John. "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods." *Advances in large margin classifiers* 10.3 (1999): 61-74.