Active Learning for Age Regression in Social Media

Jing Chen Shoushan Li Bin Dai Guodong Zhou

Natural Language Processing Lab, School of Computer Science and Technology, Soochow University, China

{jing.chen199225,chosendai}@gmail.com,
{lishoushan,gdzhou}@suda.edu.cn

Abstract. Large-scale annotated corpora are a prerequisite for developing high-performance age regression models. However, such annotated corpora are sometimes very expensive and time-consuming to obtain. In this paper, we aim to reduce the annotation effort for age regression via active learning. The key idea of our active learning approach is first to divide the whole feature space into several disjoint feature subspaces and then leverage them to learn a committee of regressors. Given the committee of regressors, we apply a query by committee (QBC) method to select unconfident samples in the unlabeled data for manual annotation. Empirical studies demonstrate the effectiveness of the proposed approach to active learning for age regression.

Keywords: Active Learning, Age Regression, Social Media.

1 Introduction

In social media, age prediction aims to determine the age of one online user by leveraging his/her published content or his/her social information. For example, Figure 1 shows an online user in a social media website. When the age is not available, we could exactly infer her age to be 23 from her published message "I'M 23". Age prediction has been an essential pre-processing step in many social applications. Generally, age classification and age regression are two foundational tasks in age prediction. Different from the age classification concerned with classifying the users into several age groups [1], age regression focuses on predicting the user's age with a discrete variable indicating an exact age number [2,3].

Conventional approaches to age regression focus on supervised learning where sufficient labeled data is essential for training the model. However, to exactly annotate the golden label of online user's age is extremely difficult [4]. A better way to obtain the labeled data is to ask for the online users to obtain their real ages. However, such way of collecting data is rather time-consuming and expensive.

In this paper, we propose an active learning approach to address above challenge by better exploiting the unlabeled data to reduce the scale of annotation data. In active lea-

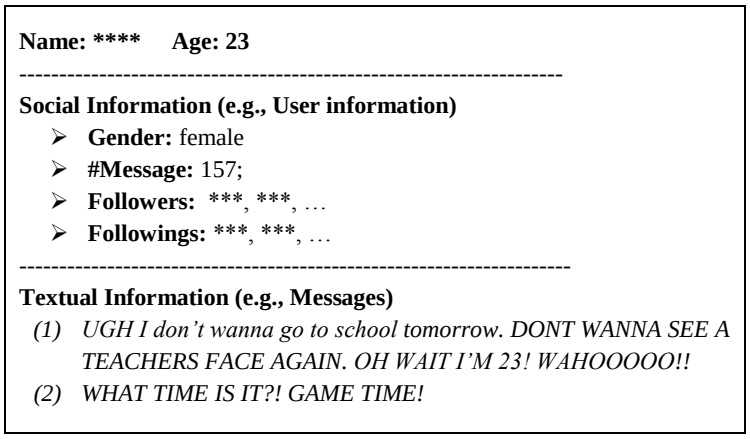

Fig. 1. A user example in a social media

rning, the simplest and most commonly used query framework is uncertainty sampling [5] where an uncertainty measurement is designed to pick most unconfidently classified instances in each iteration. In a classification problem, it is easy to measure the uncertainty by leveraging the posterior probabilities provided by the classifiers. For instance, in a 2-class classification problem, the posterior probability of one class around 0.5 is thought to be a very unconfidently classified score. However, in a regression problem, there is no such estimated probability which can be used directly because the possible predictions in regression are infinite [6].

To tackle the above difficulty, we propose a novel method to estimate the labeling unconfidence of an unlabeled instance. Specifically, our method generates multiple random feature subspaces to train a committee of regressors and then leverage the committee of regressors to estimate the labeling unconfidence of each instance. The motivation of using feature subspaces to generate multiple subspace regressors is due to the fact that the feature space (either using textual features or social features) in age regression is extremely high and we believe that a feature subspace of a certain scale is sufficient to train a good regressor. For example, in our data collected form the social website called Sina Weibo, the dimension of the textual features (i.e., word unigram features), is normally larger than 200,000 while the dimension of the social features is larger than 600,000. In principle, our method is a specific implementation of the famous Query By Committee (QBC) method which has been successfully applied in active learning [7]. For clarity, we refer to our method as subspace-based QBC. To the best of our knowledge, this is the first attempt to employ multiple feature subspaces to generate the committee in QBC for active learning in a regression task.

The remainder of this paper is organized as follows. Section 2 overviews related work on age regression and active learning for regression. Section 3 introduces some background on data collection and the basic model for age regression. Section 4 presents the active learning algorithm for age regression. Section 5 evaluates the pro-posed approach. Finally, Section 6 gives the conclusion and future work.

2 Related Work

This section gives an overview of related work on age prediction and active learning for regression respectively.

2.1 Age Prediction

Over the last decade, the overwhelming majority of studies model age prediction as a classification problem. For instance, Peersman [1] apply a text categorization approach to age classification with textual features only. Some other studies, such as Mackinnon and Warren [8], and Rosenthal and McKeown [9], explore social features to enhance the performance of age classification.

Compared to age classification, related work on age regression is much less. Nguyen [2] explore textual features, such as word unigrams, POS unigrams and bigrams, together with gender features in age regression via a linear regression model. Their empirical studies find that word unigrams can achieve reasonable performance and that POS patterns are strong indicators of the old age. Another contribution of their work is their joint model for performing age regression with three different genres of data. More recently, Nguyen [3] further explore age prediction of Twitter users with a linear regression model. They find that an automatic system can achieve better performance than human being.

To the best of our knowledge, no previous studies have been conducted their research on active learning for age prediction.

2.2 Active Learning for Age Regression

Active learning has been extensively explored in both natural language processing (NLP) and machine learning (ML) communities. For a quick and overall understanding the research issue of active learning, please refer to two comprehensive surveys, i.e., Olsson [10] and Settles [11], which discuss related studies on active learning in the NLP and ML communities respectively.

While most previous studies focus on the scenario of classification problems [12], only a few studies address the active learning issue on regression problems. Burbidge [13] employ QBC in active learning for regression by measuring disagreement as the variance among the committee members' output predictions. The committee members in their approach are generated by using several subsets of the training data.

To the best of our knowledge, no previous studies focus on active learning for regression with random feature subspaces.

3 Background

In this section, we give some background on data collection and the basic regression model for age regression.

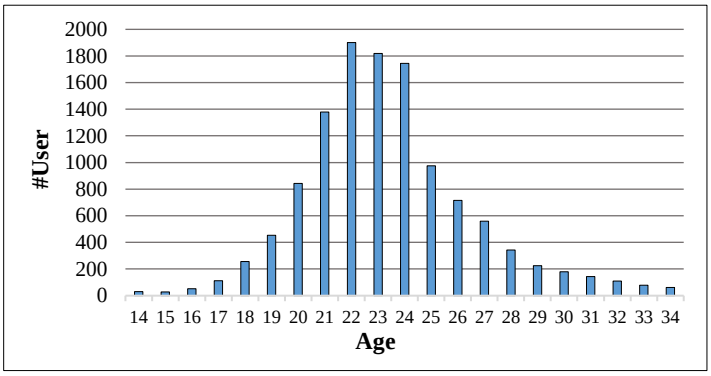

Fig. 2. User distribution in different ages

3.1 Data Collection

Our data is collected from Sina Micro-blog (http://weibo.com/), a famous Micro-blogging platform in China. From the website, we crawl each user's homepage which contains user information (e.g., name, age, gender, verified type), and their posted messages. The data collection process starts from some randomly selected users, and iteratively gets the data of their followers and followings. We remove those unsuitable users who are verified as organizations because the age attributes of these users make no sense. Besides, although the posted messages are the basic and major factor to predict user ages [2], some users post very few messages. To guarantee the reliability of the data, we remove those somehow non-active users who post less than 50 messages. In total, we collect the homepages of about 12000 users, together with their posted messages.

Figure 2 shows the user distribution in different ages. From this figure, we can see that the data distribution of user ages are rather imbalanced. Most users are young whose ages are in the range of 19-28.

3.2 Basic Model for Age Regression

In this study, we model age prediction as a regression model and apply support vector machines (SVM) to estimate the regression function [14].

Given the training data $\{x_i, y_i\}_{i=1}^{n}$ where each input $x_i \in R^d$ and the response $y_i \in R$, our goal is to find a function $f(x)$ that maps the input x_i into y_i. Suppose $f(x)$ is a linear function, taking the form:

$$f(x) = < w, x > + b \text{ with } w \in R^d,\ b \in R \tag{1}$$

where $< \cdot, \cdot >$ denotes the dot product in R^d. In $\varepsilon - SV$ regression [14], learning the function $f(x)$ becomes solving the following convex optimization problem:

Table 1. Textual and Social features in age regression

	Feature	Remarks	Examples
Textual Features	BOW	Word unigrams in the user-gene rated messages	don't, wanna, …
	POS Patterns	Top trigrams of the POS tag in user-generated messages	DT_SP_PU, PU_VV_VV,…
Social Features	Statistics	# of Messages, # of Comments , # of Followers,# of Followings	100,10,200,300
	Time	Probability distribution of the user posts messages over 24 hours (00-23)	[0.1, 0,0, …,0.2]
	Follower List	All IDs of the followers	'2919393812', '3044343944',…
	Following List	All IDs of the followings	'1976649967', '2286980683',…

$$\text{minimize} \quad \frac{1}{2}\| w \|^2 + C\sum_{i=1}^{n}(\zeta_i + \zeta_i^*) \tag{2}$$

$$\text{subject to} \quad \begin{cases} y_i - < w, x_i > -b \le \varepsilon + \xi_i \\ < w, x_i > +b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \quad \ge 0 \end{cases} \tag{3}$$

Where ζ_i, ζ_i^* are two slack variables and the constant C>0 determines the trade-off between the flatness of $f(x)$ and the amount up to which deviations larger than ε are tolerated. In age prediction, because the user ages are all integers, we round up the outputs of real numbers into integers.

4 Active Learning for Age Regression

In active learning, both labeled data L and unlabeled data U are available and the goal is to improve the performance by exploiting unlabeled data, finally reducing annotation cost. In other words, we hope to get better performance quickly when we only manually labeled a limited number of instances from unlabeled data U .

4.1 Textual and Social Features

Each user is represented by a feature vector, i.e., $x_i \in R^d$ as the input in a regression model. In the literature, various features, such as word unigrams, and social behaviors, have been successfully adopted on age prediction [9]. In this study, we categorize these features into two main groups, textual and social features. The former contains the features, generated from the user-generated messages, e.g., word unigrams, while the latter contains the features, generated from the user social behaviors, e.g., follower list and

Input:

Labeled data L;
Unlabeled pool U;

Output:

New Labeled data L

Procedure:

Loop for n iterations:

(1) Learn a committee of member classifiers using current L
(2) Use all member classifiers to label all unlabeled samples
(3) Select n most informative samples for manual annotation with the predefined unconfidence measuring strategy $unconf(x)$
(4) Move n newly-labeled samples from U to L

Fig. 3. Pool-based active learning with QBC

following list. Table 1 shows all the features in the two categories.

Among textual features, BOW features are most popular in age prediction and proven very effective due to the fact that word features reflect concerning topics, which can distinguish users of different ages. POS patterns are also popular textual features to capture the writing styles of the users.

Among social features, the 4 statistical features, i.e., those with # of, capture the social behaviors of a user. The Time features capture the user habits on posting messages. For example, users of 20-24 ages might be more likely to post their messages very late at night. Followings and followers reflect the interests of users which provide an effective window to infer users' ages.

4.2 Active Learning with QBC

Generally, active learning can be either stream-based or pool-based [15]. The main difference between the two is that the former scans through the data sequentially and selects informative samples individually, whereas the latter evaluates and ranks the entire collection before selecting most informative samples at batch. As a large collection of samples can easily gathered once in age regression, pool–based active learning is adopted in this study.

In the study, we utilize query by committee (QBC) method as our basic active learning framework. Originally, query by committee (QBC) is a group of active learning approaches which employ a committee of learners to select an unlabeled example at which their classification predictions are maximally spread [7]. Figure 3 illustrates a standard pool-based active learning algorithm with QBC method. In this algorithm, the way of learning a committee of member classifiers and the confidence measuring strategy are two crucial components which will be discussed in the next subsection in detail.

Input:

$L_{textual}$: labeled textual samples; $U_{textual}$: unlabeled textual samples

Output:

The most confidently predicted sample in $U_{textual}$

Procedure:

(1) Adopt RSG to generate *N* subspace training data { $L_{textual}^{S_1}$, $L_{textual}^{S_2}$,…, $L_{textual}^{S_N}$ }

(2) Learn *N* regression functions { $f_{tetxual}^{S_1}$, $f_{textual}^{S_2}$,…, $f_{textual}^{S_N}$ } with the obtained subspace training data.

(3) Use all regression functions to label the samples from $U_{textual}$

(4) Calculate and sort the unconfidence scores of all unlabeled samples with formula (4)

(5) Pick the sample with the maximum unconfidence score as the most unconfidently predicted sample in $U_{textual}$

Fig. 4. Unconfident samples selecting by QBC with Random Feature Subspaces

4.3 QBC with Random Feature Subspaces

To generate a committee of learners, we adopt the Random Subspace Generation (RSG) approach to generate multiple learners trained with several feature subspaces [16]. Assume $L=(x_1,x_2,...,x_n)$ the training data and x_i an *m*-dimensional vector $x_i=(w_{i1},w_{i2},...,w_{im})$, described by *m* features. RSG first randomly selects *r* ($r<m$) features and obtains an *r*-dimensional random subspace of the original m-dimensional feature space. In this way, a modified training set $L^S=(x_1^S,x_2^S,....,x_n^S)$ consisting of *r*-dimensional samples $x_i^s=(w_{i1}^s,w_{i2}^s,...,w_{ir}^s)$ $(i=1,....,n)$ is generated. Then, a subspace regression leaner can be trained in random subspaces x^S using the modified training set. In our implementation, we set *N* to be m/r and thus *N* disjoint feature subspaces are utilized to generate *N* subspace regression learners.

Active learning aims to select the most uncertain (unconfident) sample rather than the most certain sample. Thus, we select an unlabeled example at which their regression predictions are maximally disagreed. Formally, given the regression results from the committee of learners $(y_1^{'},y_2^{'},...,y_N^{'})$, the unconfidence score is calculated as follows :

$$unconf(x)=\log\left(\sum_{i=1}^{N}(y^{'}-y_i^{'})^2\right) \tag{4}$$

Where $y^{'}$ is the estimated result of the committee, calculated as follows:

$$y^{'}=\frac{1}{q}\sum_{i=1}^{q}y_i^{'} \tag{5}$$

The more the unconfidence score is, the more unconfidently the sample is predicted.

Figure 4 shows the algorithm of our QBC-based approach to selecting unconfident samples. Note that we only give the algorithm description on the textual features. A similar description is obvious for the social features and joint features.

5 Experimentation

In this section, we have systematically evaluated our approach to active learning for age regression.

5.1 Experimental Settings

Data Setting

The data collection has been introduced in Section 3.1. We extract a balanced data set from the collected data by selecting 200 samples in each age and the age is limited in the range of 19 to 28, totally 10 age categories. We use 80% of the data in each age category as the training data and the remaining 20% data as test data. In active learning, we randomly select 10 users in each age category from the training data as the initial labeled data and the remaining training data as unlabeled data.

Regression Algorithm & Features

We use the libSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) tool to implement our SVM regression algorithm with the linear kernel and the features as described in Table 1.

Evaluation Metric

We employ the coefficient of determination R^2 to measure the regression performance. Coefficient of determination R^2 is used in the context of statistical models with the main purpose to predict the future outcomes on the basis of other related information. R^2 is a number between 0 and 1. R^2 nearing 1.0 indicates that a regression line fits the data well [17].

5.2 Experimental Results

In this subsection, we present the experimental results when we leverage different kinds of features respectively including textual features, social features and joint features (combine social and textual features) to perform active learning algorithm on age regression. What's more, we adopt several comparable experiments during the process of investigating the effect of active learning on age regression.

Before reporting the results of active learning, we first investigate the performances of different kinds of features for age regression in a supervised learning setting. Table 2 shows the age regression results of fully supervised learning (i.e., all training data is used as labeled data to train the regressor) when different kinds of features are utilized. From this table, we can see that BOW, following list and follower list occupy a large amount of features and perform apparently better than other kinds of textual and social features. We can also see that adding POS features in textual features is not helpful, while adding other types of features in social features is more helpful than using following list features only. Finally, we can see that the performance becomes best at 0.520 in R^2, when both textual and social features are employed.

Table 2. Performance of fully supervised learning when different kinds of features are used

	Feature	# of features	R^2
Textual Features	BOW	116463	0.413
	Patterns POS	10153	0.087
	ALL	126616	0.409
Social Features	Statistics	4	0.010
	Time	24	0.013
	Follower List	103080	0.168
	Following List	80904	0.256
	ALL	184012	0.269
Joint Features	Textual+Social	310628	**0.520**

For thorough comparison, some active learning approaches are implemented including:

- **Random:** which randomly selects the samples from the unlabeled data for manual annotation.
- **Subsample-based QBC:** which divides labeled samples into several groups and utilize these groups to train several age regressors as the committee of member regressors which are then utilized in a QBC active learning algorithm. In our implementation, we change the numbers of groups from 5 to 20 and find that the performances remain similar. The reported results are obtained when the group number is 10. This is the approach proposed by Burbidge [13] for a regression task.
- **Subspace-based QBC** (Our approach)**:** which concretes the implementation in Section 4.3.

In our implementation, we run these approaches 5 times and report the average results eventually. The number of the feature subspaces is set to be 8 when the social features, textual features or joint features are utilized individually.

Figure 5 compares our approach with other active learning approaches by varying the number of the selected samples for manually annotation and all approaches are performed with social features, textual features or joint features respectively. From this figure, we can see that **Subsample-based QBC** is effective when social features are used. However, when textual and joint features are used, Subsample-based QBC performs even worse than the random selection approach. Our approach **Subspace-based QBC** apparently outperforms other two approaches no matter what kind of features is used. Significance test with *t*-test shows that our approach significantly outperforms the other two approaches (p-value<0.01) when less than 600 unlabeled samples are selected no matter of features are used.

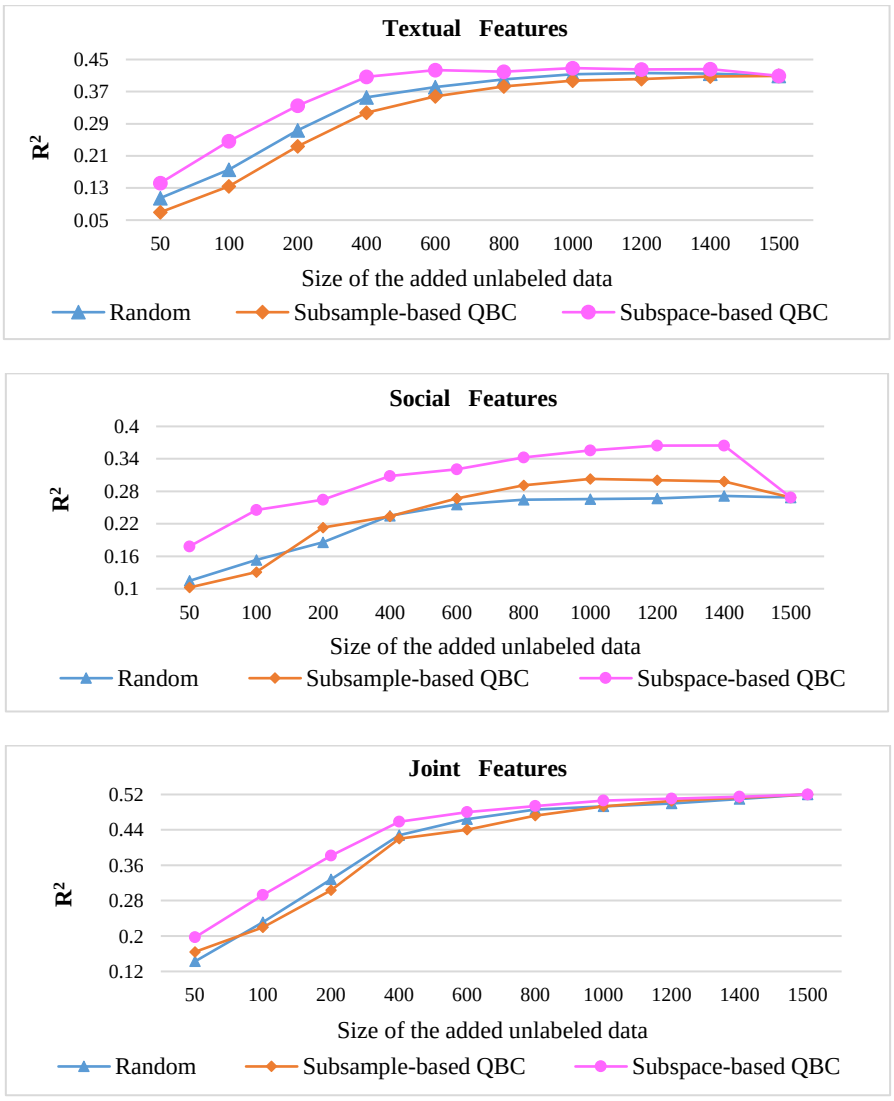

Fig. 5. Performance comparison of active learning approaches with different features

The number of the feature subspace is an important parameter in our approach. Figure 6 shows the performance of QBC based on subspace with different size of the feature spaces when we utilize textual features, social features or joint features individually. From Figure 6, we can see that our approach Subspace-based QBC consistently outperforms the random selection approach when varying the number of feature subspaces. For a nice performance, a choice of the number between 6 and 16 is recommended to be the size of feature subspace.

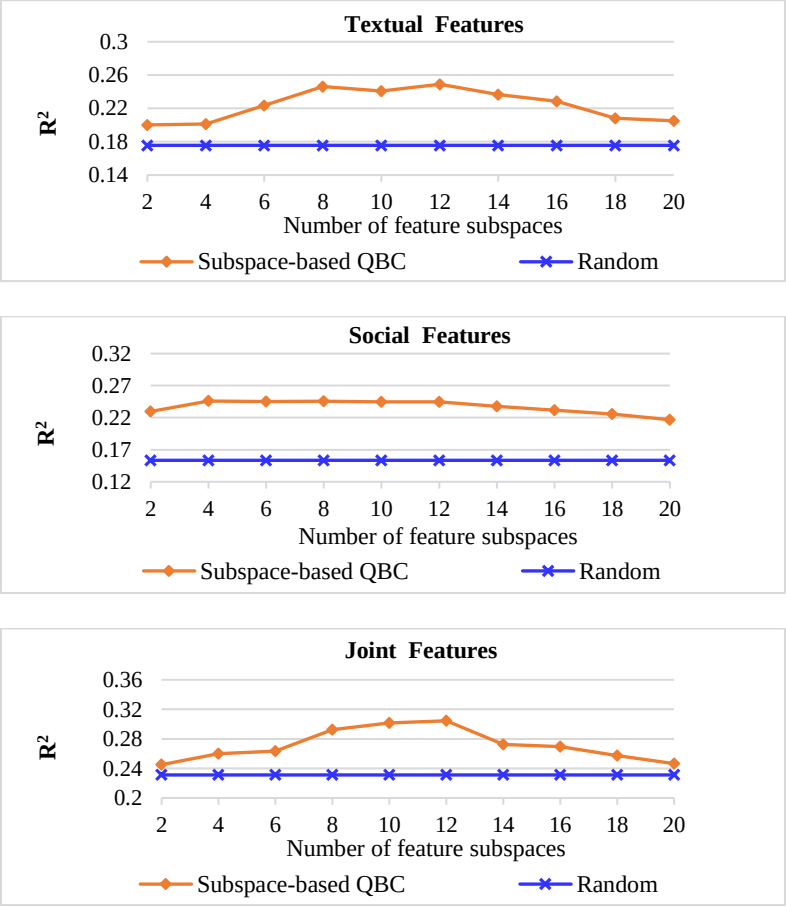

Fig. 6. Performance of Subspace-based QBC over varying sizes of feature subspaces when using textual features, social features, and joint features respectively.

6 Conclusion

In this paper, we propose an active learning approach to age regression for better exploiting the unlabeled data to improve the performance. Our approach leverages three kinds of features, namely textual features, social features and joint features (combining textual and social features). Moreover, we propose a QBC-style approach to active learning for age regression. In our approach, we solve the unconfidence estimation problem in our regression model by using a committee of feature-subspace regressors. Evaluation shows that our approach, namely subspace-based QBC, effectively improves the performance in active learning.

In our future work, we would like to improve the performance on age regression by exploring more features. Moreover, we would like to apply our approach to active learning on regression in some other NLP tasks.

Acknowledgments

This research work has been partially supported by two NSFC grants, No.61375073 and No.61273320, one the State Key Program of National Natural Science of China No.61331011.

References

1. Peersman, C., Daelemans, W., Vaerenbergh, L.V.: Predicting Age and Gender in Online Social Networks. In: Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents SMUC, pp. 37-44 (2011)
2. Nguyen, D., Smith, N.A., Rose, C.P.: Author Age Prediction from Text using Liner Regression. In: Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pp. 115-123 (2011)
3. Nguyen, D., Gravel, R., Trieschnigg, D., Meder, T.: "How Old Do You Think I Am?" : A Study of Language and Age in Twitter. In: Proceedings of AAAI Conference on Weblogs and Social Media, pp. 439-448 (2013)
4. Nguyen, D., Trieschnigg, D., Dogruöz, A.S., Gravel, R.: Why Gender and Age Prediction from Tweets is Hard: Lessons from a Crowdsourcing Experiment. In: Proceedings of COLING, pp. 1950-1961 (2014)
5. Lewis, D.D., Catlett J.: Heterogeneous Uncertainty Sampling for Supervised Learning. In: Proceedings of the International Conference on Machine Learning, pp. 148-156 (1996)
6. Zhou, Z.H., Li, M.: Semi–supervised Regression with Co-Training. In: Proceedings of IJCAI, pp. 908-913 (2005)
7. Freund, Y., Seung, H.S., Shamir E., Tishby, N.: Selecting Sampling Using the Query by Committee Algorithm. Machine Learning 28(2-3), 133-168 (2001)
8. Mackinnon, I., Warren, R.: Age and Geographic Inferences of the Live Journal Social Network. In: Proceedings of ICML, pp.176-178 (2006)
9. Rosenthal, S.: Age Prediction in Blogs: A Study of Style, Content, and Online Behavior in Pre- and Post-Social Media Generations. In: Proceedings of ACL, pp. 763-772 (2011)
10. Olsson, F.: A Literature Survey of Active Learning Machine Learning in the Context of Natural Language Processing. SICS Technical Report (2009)
11. Settles, B.: Active Learning Literature Survey. Computer Sciences Technical Report 1648 39(2), pp. 127–131 (2010)
12. Li, S.S., Xue, Y.X., Wang, Z.Q., Zhou, G.D.: Active Learning for Cross-domain Sentiment Classification. In: Proceedings of IJCAI, pp. 2127-2133 (2013)
13. Burbidge, R., Rowland J.J., King R.D.: Active Learning for Regression based on Query by Committee. In: Proceedings of Intelligent Data Engineering and Automated Learning (IDEAL), pp. 209-218 (2007)
14. Vapnik, V.N.: The Nature of Statistical Learning Theory, pp. 988-999 (1995)
15. Sassano, M.: An Empirical Study of Active Learning with Support Vector Machines for Japanese Word Segmentation. In: Proceedings of meeting of the Association for Computational Linguistics, pp. 505-512 (2002)
16. Ho, T.: The Random Subspace Method for Constructing Decision Forests. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(8), pp. 832-844 (1998)
17. Cameron, A., Windmeijer, F.: R-squared Measures for Count Data Regression Models with Applications to Health-care Utilization. Journal of Business and Economic Statistics 14(2), 209-220 (1993)