Conscientiousness Measurement from Weibo's Public Information

Dong Nie, Lin Li, and Tingshao Zhu*

{Institute of Psychology, University of Chinese Academy of Sciences}, CAS, Beijing, 100190, China
ginobilinie@gmail.com
http://ccpl.psych.ac.cn

Abstract. We apply a graph-based semi-supervised learning algorithm to identify the conscientiousness of Weibo users. Given a set of Weibo users' public information(e.g., number of followers) and a few labeled Weibo users, the task is to predict conscientiousness assessment for numeric unlabeled Weibo users. Singular value decomposition(SVD) technique is taken for feature reduction, and K nearest neighbor(KNN) method is used to recover a sparse graph. The local and global consistency algorithm is followed to deal with our data. Experiments demonstrate the advantage of semi-supervised learning over standard supervised learning when limited labeled data are available.

Keywords: graph-based semi-supervised learning, conscientiousness identification, KNN, SVD

1 Introduction

Personality can be defined as a set of characteristics which make a person unique, and the study of personality is of central importance in psychology. Among personality related researches, Big-Five theory is the mostly used one, it proposes five basic traits to form human personality: openness, conscientiousness, extraversion, agreeableness and neuroticism [10]. Conscientiousness, as one trait of Big-Five theory, is the state of being thorough, careful, or vigilant. Conscientious individuals are generally hard working and reliable. When taken to an extreme, they may also be "workaholics", perfectionists, and compulsive in their behavior. People who are low on conscientiousness are not necessarily lazy or immoral, but they tend to be more laid back, less goal-oriented, and less driven by success [16]. Conscientiousness measurement are mostly through self-report [27], which is time-consuming and sometimes,subjective. Some researchers made efforts to others-report [13, 6, 9], however, it is a big problem to achieve enough labeled data due to not only time-consuming and expensive, but also privacy problems.

Sina Weibo is now one of the most popular service in mainland China, it has already attracted more than 300 million user to register the service [23], and many people spend much time on the service, thus, researchers say it has become an important part of user's life [7]. Many researches were done to found out the relationship between microblog usage and user's personality [11, 20, 13].

* Corresponding Author: Tingshao Zhu,tszhu@psych.ac.cn.

In recent years, there has been a substantial amount of work exploring how to incorporate unlabeled data into supervised learning, and several semi-supervised learning approaches have been proposed [4, 31, 28, 3, 30]. Successful applications have been made in many areas, such as computer vision [2, 29], and information retrieval [24]. Semi-supervised learning has also been used in the context of microblog classification [17]. In many scenes, semi-supervised learning algorithms outperforms standard supervised learning algorithms.

In this paper, we propose a graph-based semi-supervised learning approach [28] to the problem of conscientiousness measurement. We randomly collected 562 Weibo users' public information using Sina Weibo API, and retrieved corresponding conscientiousness extent through self-report method. Then local and global consistency method was used to combine the labeled and unlabeled data.

The rest of this paper is organized as follows. In Section 2, we introduce some related work. Section 3 will describe the dataset in detail. Section 4 will talk about the graphs. In Section 5, we present the detailed algorithm. Experiment results will be discussed in Section 6. At last, we will give a conclusion in Section 7.

2 Related Work

Personality analysis based on social media has received considerable attention recently [11, 1, 20, 13]. They mainly collected internet data and corresponding labeled data, and then applied supervised learning approaches, such as, classification and regression analysis, to build the mode. However, the problem is that training data is always scarce, meanwhile large scale of unlabeled data is often easy to retrieve.

Many methods have been proposed to deal with the problem. Among them, semi-supervised learning algorithms received great attention in the past few years, since they could perfectly make use of unlabeled data. Generative models are used for text classification [24] with both labeled and unlabeled data, but the assumption is that the data should obey a certain distribution. Co-training algorithm has been proposed in [5], much research has been done, but it requires features can be spilt into two conditionally independent sets. Transductive support vector machines(TSVMs) have also been investigated by [19, 8]. These years, many computer scientists are devoted to graph-based semi-supervised learning methods and propose a series of effective algorithms [4, 31, 28, 3], and these algorithms have been widely used in many fields.

In this paper, local and global consistency(LGC) method [28] are integrated, because people with similar Weibo data is supposed to have the same personality. We gathered a number of Sina Weibo users' public information, and asked part of them to finish a big-five inventory [21] online. We only focus on one trait:conscientiousness in the work. We used singular value decomposition(SVD) to select features, and added a graph sparsification step to optimize the graph. The LGC semi-supervised learning algorithm was taken to give a ternary classification over the collected data, and thus users' conscientiousness extents were assessed.

3 Dataset

The dataset consists of 562 copies of Sina Weibo users' information together with corresponding conscientiousness scores and much more non-labeled data. The Weibo data is about the user items on Weibo service, and it can be spilt into several categories as follows:

1. information in user's personal profile, including nickname, address, gender, birthday, personalized domain name, description and so on.
2. information about friends and followers, for example, the number of friends.
3. statistical information for statuses, such as average count of statuses per day and proportion of originality statuses.
4. basic setting information, for example, whether to allow all to comment.
5. tags information
6. trends information
7. others

We didn't focus much on users' behavior information in the work, for example, user's specific Weibo statuses. The conscientiousness scores are measured in continuous value. Since only a few users are labeled, and more users' Weibo data are available, it is very suitable to use semi-supervised learning techniques for our problem.

3.1 Data Collection

Using Sina Weibo API [1] , we first collected 999,9999 Weibo user IDs, then randomly chosen 10,000 user IDs. Using Sina Weibo API again, we crawled 10,000 users' Weibo data which has been described above from these 10,000 Weibo IDs. In this way, we successfully captured a 10,000-Weibo-user dataset. Thirdly, using Weibo service's @ function, we invited the users to be volunteers to finish big-five inventory online. At last, we collected 562 copies of qualified questionnaires. Hence, we had 562 copies of conscientiousness extent data. The whole process took over one month, and volunteers had got reimbursement in return.

The collected Weibo-user dataset(562 copies of labeled Weibo data) is the basis of our experiment.

3.2 Feature Extraction

As our collected Weibo dataset is relatively simple(not contain much behavior data), the preprocessing work is straight forward. For some features, we followed the original data directly, for example, statistical information about statuses. For others, we used simple statistical methods to deal with data to extract features. We had totally extracted 45 features for a user from the Weibo data. Some of the features are listed in Table 1.

After feature extraction, we normalize the Weibo data to make the data equally measured as follows.

$$x = (x - MinValue)/(MaxValue - MinValue)$$

[1] `http://open.weibo.com`

Table 1. Part of Extracted Features

Feature	Description
allow_all_comment	Whether to allow all users to freely comment
bi_all_followers_count	The number of user's followers
bi_all_friends_count	The number of user's friends
description	The length of user's description
statuses_weibo_count	The number of user's statuses
original_rate	The rate of original statuses
screen_name_length	The length of screen name
users_tags_count_100	The number of tags whose popularity is less than 100
users_tags_count_100_1000	The number of tags whose popularity is between 100 and 10,000
users_tags_count_100_1000	The number of tags whose popularity is over 10,000
first_weibo_time	The average time to give first status per day
last_weibo_time	The average time to give last status per day
verified	Whether the user's Weibo account has been verified
...	...

where x is the value of a dimension for a user, while $MinValue$ and $MaxValue$ respectively represent the maximum and the minimum value of this feature dimension for all users.

3.3 Conscientiousness Scores Discretization

As we received 562 copies of effective big-five personality questionnaires, we paid attention to one trait:conscientiousness only in the work. We calculated the conscientiousness score for each user according to their questionnaire based on corresponding rule, however, it is measured in continuous value, and it should be discretized.

In various personality researches [11, 20, 26], level of grouping method is often used to discretize continuous value of personality trait. Specifically speaking, we first calculate the mean(μ) and standard deviation(σ) for the sample, and then subjects whose conscientiousness scores are greater than $(\mu + \sigma)$ are grouped to be high level, while subjects whose conscientiousness scores are lower than $(\mu - \sigma)$ are grouped to be low level, and subjects whose scores are between $(\mu - \sigma)$ and $(\mu + \sigma)$ are grouped to be normal level. According to the definition of conscientiousness, individuals with high level mean they are conscientious, even workaholics to some degree, on the contrary, individuals with low level mean they tend to be more laid back, less goal-oriented, and less driven by success, individuals with middle level are supposed to be normal. To have a simple description later, we abbreviate the three levels as "conscientious", "immoral" and "normal" respectively. Therefore, the conscientiousness label set C can be represented by these three levels:

$$C = \{\text{“}conscientious\text{”}, \text{“}immoral\text{”}, \text{“}normal\text{”}\}$$

4 The Graphs

The semi-supervised conscientiousness measurement problem is described as follows. There are 562 Weibo users $x_1, x_2...x_{562}$, each represented by a set of features discussed above. We randomly choose $l \leq 562$ Weibo users from the labeled dataset, and suppose the l Weibo users to be labeled with $y_1, y_2...y_l \in C$ respectively. The remaining data is set to be unlabeled. The goal is to predict the categories of the unlabeled points using method from [28].

4.1 Feature Reduction

Usually, multidimensional data may be represented approximately in fewer dimensions due to redundancies in data, which may improve the prediction accuracy [25]. Since the original feature space has 45 dimensions, we attempt to take singular value decomposition(SVD) method [12], which is a well known matrix factorization technique, to reduce dimensionality of feature space [15, 22]. We simply describes the SVD methods as follows.

Suppose $\underset{n\times 45}{A}$ be our original data space, then use SVD technique to factor A into three matrices:

$$\underset{n\times 45}{A} = \underset{n\times r}{U} \sum_{r\times r} \underset{r\times 45}{V}$$

where, matrix $\sum$ is a diagonal matrix containing the singular values of the matrix A, here are exactly r singular values, where r is the rank of matrix A. The rank of a matrix is the number of linearly independent rows or columns in the matrix, and it means independent information in our data space here.

To accomplish we can simply keep the first k singular values in $\sum$, where $k \leq r$. This will give us the best rank-k approximation to original data space A, and thus has effectively reduced the dimensionality of our original space. In our experiment, the dimensionality of original feature space is reduced to 34.

4.2 Graph Construction

We first compute measure similarity between Weibo user x_1 and user x_2 by features, and a larger similarity implies that the two users have more chances to be the same conscientiousness extent. Details can be found in Section 5.

An undirected graph $G = (V, E)$ is formalized with n nodes V, and weighted edges E among the nodes. Each Weibo user is a node in the graph, including the unlabeled users. The node of labeled user is also labeled with conscientiousness extent in the graph. Each Weibo user is connected to any other users by similarity computed between the two users, no matter the user is labeled or not. Then a fully connected graph is constructed.

4.3 Graph Sparsification

As described above, the graph for our semi-supervised learning problem is a fully connected one. To ensure that the semi-supervised learning algorithm remain efficient and robust to noise, a sparse weighted subgraph from the fully connected graph is needed. There are few researches on graph construction [18], though graph-based semi-supervised learning has received much attention recently. K Nearest Neighbors (KNN) is the most common used algorithm to recover a sparse subgraph. Roughly speaking, each node merely connects to its k nearest neighbors to form a subgraph. Specifically, for each point in the fully connected graph, using similarities in the fully connected graph, searches for the k nearest points to it without considering itself. Thus we can recover a sparse subgraph with this method.

5 Algorithms

We use the simple Local and Global Consistency(LGC) algorithm [28] on the Weibo dataset, the following formula depicts the essence of this algorithm:

$$\min_f \{\sum_{i=1}^{l} (f(x_i) - y_i)^2 + f^T \Delta f\}$$

where $f(x)$ is the decision function, y is the label for each node and Δ is the graph combinatorial Laplacian. The LGC method allows $f(x_l)$ to be different from y_l with penalty term, in other words, this method can accommodate noise.

We choose the following function to compute similarity between two users:

$$W_{ij} = \exp\left(-d(x_i, x_j)/2\sigma^2\right)$$

, where

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{t} (x_{ik} - x_{jk})^2}$$

, where t is the dimension of the feature space.

We now describe the LGC algorithm in detail:

1.Form the affinity matrix W using the function discussed above.

2.Recover a sparse affinity matrix using KNN method.

3.Construct the matrix $S = D^{-1/2} W D^{-1/2}$, where $D = diag(\sum_j w_{ij})$.

4.Iterate $F(t+1) = \alpha S F(t) + (1-\alpha) Y$ until convergence, where $\alpha \in (0, 1)$.

5.Let F^* denote the limit of the sequence $\{F(t)\}$. Label each point x_i as a label $y_i = \arg\max_{j \le c} {F_{ij}}^*$.

The above algorithm has been proved convergent, and we can computer F^* without iterations:

$$F^* = (I - \alpha S)^{-1} Y$$

where I is a identity matrix.

The LGC method described above solves a set of linear equations so that the predicted label of each example is the result of considering local and global consistency.

The algorithm makes it that nearby points are likely to have the same label and points on the same structure are likely to have the same label, too. This algorithm successfully makes use of a amount of labeled and unlabeled points to build a classification, moreover it doesn't limit to binary classification problems, therefore, this method is suitable for our Weibo dataset.

6 Experiment results

As it is very difficult to collect labels for the entire 10,000-Weibo-user dataset, we can only conduct experiments on the 562-Weibo-user dataset.

We evaluated LGC method on the Weibo user conscientiousness measurement tasks. For each task, we gradually increased the labeled set size systematically, performed 10 random trials for the labeled set size. In each trial we randomly sampled a labeled set with the specified 562-Weibo-user dataset, if a class was missing from the sampled labeled set, we redid the random sampling, and the remaining data were used as the unlabeled set. We select the parameter combination by a grid search with the three parameters (k, σ, α), and the results are as follows. The k in KNN sparsification is set to 4, the parameter in the weight function is selected as $\sigma = 2$, and the iteration coefficient(α) is set to 0.2.

We report the classification accuracies with LGC method on two graphs: fully connected graph and KNN sparse graph. And we also compare two feature space: original and compressed feature space. To compare the graph-based semi-supervised learning algorithm against a standard supervised learning algorithm, we choose one-vs-rest SVMs as baseline method, and we use SVM-Light(svmlight.joachims.org/) as the tool to implement our ternary classification problems. Specifically, we create three binary classifications, one for each class against all the other classes, and select the class with the largest margin. We choose RBF kernel for the SVM classifications, and the width of RBF kernel is set to 1.4. The results are presented in Figure 1.

"SVD" means using Singular Value Decomposition method to deal with original feature space, "sparse" indicates using KNN methods to recover a sparse subgraph, and "original" denotes original feature space and fully connected graph. As shown in Figure 1, we find that LGC method outperforms the RBF kernel SVM baseline a lot, especially when the labeled set size is small. LGC method with SVD processing and KNN sparsification performs best over all methods. When the labeled data set comes to 50, the accuracy can approximate 80%, which is a good performance in conscientiousness measurement. The accuracy can be improved to a certain extent if we use KNN method to construct a sparse subgraph. Using SVD method to compress feature space can also improve the classification accuracy in a small degree, it may be because manifold learning is not the best choice in semi-supervised learning for truly high dimensional data [14].

7 Conclusions

In this paper, we have investigated a local and global consistency based semi-supervised learning algorithm for conscientiousness measurement, which conforms two assump-

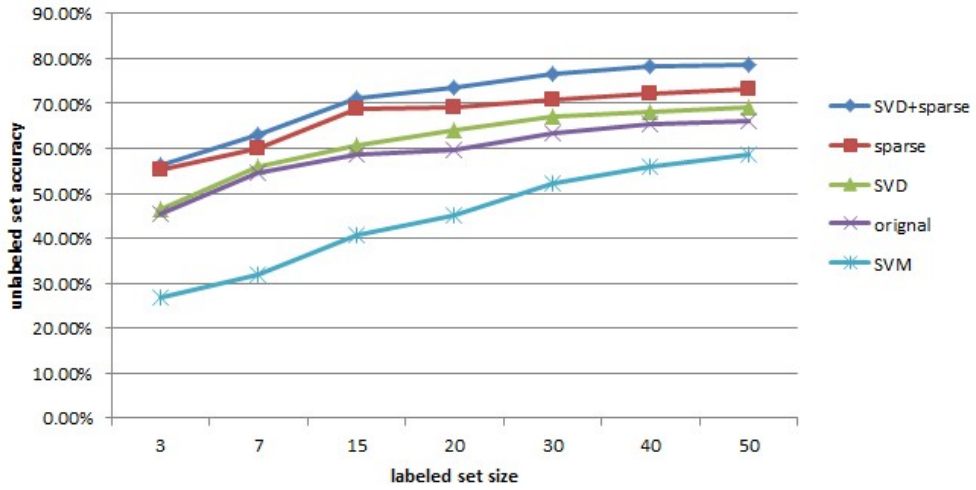

Fig. 1. The Accuracy of LGC and SVM.

tions: similar examples should have similar labels and examples in similar structure should have similar labels. The algorithm makes full use of both labeled data and unlabeled data.

In our research, we obtained a set of Sina Weibo data, with 562 conscientiousness labels. We conducted our experiments over the dataset. Singular Value Decomposition(SVD) method was used to perform dimensionality reduction and K nearest neighbor(KNN) method was used to recover a sparse subgraph. Then the local and global consistency(LGC) algorithm was taken to give a ternary classification over the dataset. Our experiment shows that LGC algorithm achieves better performance when only very few labeled examples are available.

Apparently, there exists vast space we can do to promote the performance. Next, we will focus on model selection and parameter selection to better construct the graph. Meanwhile, we will try to incorporate more background knowledge into the learning process. In the future, we will use semi-supervised learning method to predict personality traits on large dataset.

Acknowledgments

The authors gratefully acknowledge the generous support from National High-tech R&D Program of China (2013AA01A606), NSFC (61070115), Institute of Psychology (113000C037), Strategic Priority Research Program (XDA06030800) and 100-Talent Project (Y2CX093006) from Chinese Academy of Sciences.

References

1. Jun Hakura Atsunori Minamikawa, Hamido Fujita and Masaki Kurematsu. Personality estimation application for social media. In *Frontiers in Artificial Intelligence and Applications*, volume 246, 2012.
2. Balcan, M.-F, Blum.A, Choi, P. P.and Lafferty, J.Pantano.B, Rwebangira, M. R., and X Zhu. Person identification in webcam images: An application of semi-supervised learning. In *ICML 2005 Workshop on Learning with Partially Classified Training Data*, 2005.
3. Niyogi P.and Sindhwani V Belkin, M. On manifold regularization. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, 2005.
4. A. Blum and S Chawla. Learning from labeled and unlabeled data using graph mincuts. In *Proc. 18th International Conf. on Machine Learning*, 2001.
5. and Mitchell T Blum, A. Combining labeled and unlabeled data with co-training. In *COLT: Proceedings of the Workshop on Computational Learning Theory*, 1998.
6. Tom Buchanan and John L. Smith. Using the internet for psychological research: Personality testing on the world wide web. *British Journal of Psychology*, 90(1):125–144, 1999.
7. Belinda Cao. Sina's weibo outlook buoys internet stock gains: China overnight. Technical report, Bloomberg, 2012.
8. Sindhwani V. Chapelle, O. and S. S Keerthi. Branch and bound for semisupervised support vector machines. In *Advances in Neural Information Processing Systems (NIPS)*, 2006.
9. Byers A Chris Sumner, M.S. Determining personality traits and privacy concerns from facebook activity. *Black Hat Briefings*, 11, 2011.
10. D.C Funder. Personality. *Annu. Rev. Psychol.*, 52:197–221, 2001.
11. J. Golbeck, C. Robles, and K. Turner. Prediciting personality with social media. In *Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems*, pages 253–262. ACM, 2011.
12. G. H. Golub and C. Reinsch. Singular value decomposition and least squares solutions. *Numerische Mathematik*, 14:403–420, 1970.
13. S. Gosling, A. Augustine, S. Vazire, N. Holtzman, and S. Gaddis. Manifestations of personality in online social networks: Self-reported facebook related behaviors and observable prole information. *Cyberpsychology behavior and social networking*, 14:483–488, 2011.
14. Y. Grandvalet and Y Bengio. Semi-supervised learning by entropy minimization. In Y. Weiss In L. K. Saul and L. Bottou, editors, *Advances in neural information processing systems 17*. Cambridge, MA: MIT Press., 2005.
15. Y. Kamp H. Bourlard. Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, 59:291–294, 1988.
16. Ones Hogan, Joyce and Deniz S. *Handbook of personality psychology*, chapter Conscientiousness and integrity at work, pages 849–870. San Diego, CA, US: Academic Press, 1997.
17. Naoki Yoshinaga Hongguang Zheng, Nobuhiro Kaji and Masashi Toyoda. A study on microblog classification based on information publicness. In *DEIM Forum*, 2012.
18. Tony Jebara, Jun Wang, and Shih-Fu Chang. Graph construction and b-matching for semi-supervised learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 441–448, New York, NY, USA, 2009. ACM.
19. T. Joachims. Transductive inference for text classification using support vector machines. In *Proc. 16th International Conf. on Machine Learning*, pages 200–209, Morgan Kaufmann, San Francisco, CA., 1999.
20. Jonathan Ramsay Lin Qiu, Han Lin and Fang Yang. You are what you tweet: Personality expression and perception on twitter. *Journal of Research in Personality*, 46:710–718, December 2012.
21. Deary I. Whiteman M Matthews, G. *Personality Traits*. Cambridge University Press, 2006.

22. Luis M. Rocha Michael E. Wall, Andreas Rechtsteiner. Singular value decomposition and principal component analysis. *A Practical Approach to Microarray Data Analysis*, pages 91–109, 2003.
23. Steven Millward. China's forgotten 3rd twitter clone hits 260 million users. Technical report, techinasia.com, 2012-10-22.
24. McCallum A. K. Thrun S. Nigam, K. and T. M Mitchell. Learning to classify text from labeled and unlabeled documents. In *AAAI-98, 15th Conference of the American Association for Artificial Intelligence*, pages 792–799, 1998.
25. G.W. Furnas T.K. Landauer S. Deerwester, S.T. Dumais and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 1990.
26. Willi H. Wiesner Susan L. Kichuk. The big five personality factors and team performance: implications for selecting successful product design teams. *Journal of Engineering and Technology Management*, 14:195–221, 1997.
27. E.R Thompson. Development and validation of an international english big-five mini-markers. *Personality and Individual Differences*, 45(6):542–548, 2008.
28. Bousquet O. Lal T. Weston J. Zhou, D. and B Schlkopf. Learning with local and global consistency. In *Advances in Neural Information Processing System*, 2004.
29. Chen K.-J. Zhou, Z.-H. and H.-B Dai. Enhancing relevance feedback in image retrieval using unlabeled data. *ACM Transactions on Information Systems*, 24:219–244, 2006.
30. Z.-H. Zhou and M Li. Semi-supervised regression with co-training. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2005.
31. Ghahramani Z. Zhu, X. and J Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *The 20th International Conference on Machine Learning (ICML)*, 2003.