4-2016

Personal credit profiling via latent user behavior dimensions on social media

Guangming GUO

Feida ZHU
Singapore Management University, fdzhu@smu.edu.sg

Enhong CHEN

Le WU

Qi LIU

See next page for additional authors

Personal Credit Profiling via Latent User Behavior Dimensions on Social Media

Guangming Guo[1,2], Feida Zhu[2], Enhong Chen[1(✉)], Le Wu[1], Qi Liu[1], Yingling Liu[1], and Minghui Qiu[2]

[1] School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China
[2] School of Information Systems, Singapore Management University, Singapore, Singapore
guogg@mail.ustc.edu.cn

Abstract. Consumer credit scoring and credit risk management have been the core research problem in financial industry for decades. In this paper, we target at inferring this particular user attribute called credit, i.e., whether a user is of the good credit class or not, from online social data. However, existing credit scoring methods, mainly relying on financial data, face severe challenges when tackling the heterogeneous social data. Moreover, social data only contains extremely weak signals about users' credit label. To that end, we put forward a Latent User Behavior Dimension based Credit Model (LUBD-CM) to capture these small signals for personal credit profiling. LUBD-CM learns users' hidden behavior habits and topic distributions simultaneously, and represents each user at a much finer granularity. Specifically, we take a real-world Sina Weibo dataset as the testbed for personal credit profiling evaluation. Experiments conducted on the dataset demonstrate the effectiveness of our approach: (1) User credit label can be predicted using LUBD-CM with a considerable performance improvement over state-of-the-art baselines; (2) The latent behavior dimensions have very good interpretability in personal credit profiling.

1 Introduction

Accurate assessment of consumers' credit risk has a profound impact on P2P lending's success. Traditional consumer credit scoring literatures have proposed various statistical methods for credit risk management [4]. Advanced methods using data mining approach [24] and machine learning approach [14] have also been proposed in recent years. Mostly, the employed consumer data for credit analysis in these studies is composed of historical loan/payment records, credit reports or demographic information like salary and education. However, according to American Consumer Financial Protection Bureau[1], almost one in ten American consumers has no credit history until 2015, not to mention other less developed countries. Even for users with credit history, online P2P lending

[1] http://files.consumerfinance.gov/f/201505_cfpb_data-point-credit-invisibles.pdf.

J. Bailey et al. (Eds.): PAKDD 2016, Part II, LNAI 9652, pp. 130–142, 2016.
DOI: 10.1007/978-3-319-31750-2_11

companies can't access their financial transaction data freely, which is usually dispersed among various institutions and companies. To make it worse, demographic survey data usually costs a lot to collect and validate, which cannot be afforded by these small loan companies.

In the era of social media, the situation is changing. The ever-growing online micro-blogging services have become indispensible for our everyday lives. Most of us rely on social media to share, communicate, discover and network [12]. Meanwhile, tons of **U**ser **G**enerated **C**ontent (UGC), such as status updates, retweets, replies etc., becomes available on social media. The practice of harnessing this personal UGC on social media for credit profiling, becomes more and more prevalent with the blossoming of online Internet finance startups like Kabbage[2] and ZestFinance[3]. For individuals applying for small loans, the online social data provides great opportunities to investigate their credit risks with unprecedented data scale, coverage, granularity and nearly no cost while preserving their privacy.

However, social data, especially the tweet data, is inherently heterogeneous, dynamic and even noisy. For instance, users on social media frequently invent new words to express their feelings and thoughts. Different from financial data or survey data, tweets are usually informal and fragmented since they are limited to be 140-character-long and diverse in topics [15]. What's more, user credit is a particularly private attribute, even more sensitive than age or gender in most cases. Users seldom generate credit related personal data on the social web. Consequently, social data only contains extremely weak signals about user credit risk. Primary experiment results show that the best prediction accuracy we can achieve is only 57.2 % with thousands of manually defined social features as input. All the above facts pose great challenges for us to leverage the social data for personal credit profiling, i.e., assessing one's credit risk into classes of "good" or "bad" [9] from social data. To our surprise, we find that some kinds of behaviors extracted from tweets, such as posting time of tweets, is informative for credit profiling (Cf.Sect. 2 for details). Unlike tweet content, behavior data is usually precise and formal, and reflects users' behavior habits and characters more comprehensively and directly. This observation is a good example of the old view that characters or habits are also key factors affecting people's credit risk. As far as we know, existing user profiling techniques only treat behavior data as an additional feature source [22], couldn't extract users' habits and characters from it for credit profiling very well.

To achieve the goal of personal credit profiling on social media, we propose the Latent User Behavior Dimension based Credit Model (LUBD-CM), which explicitly models users' behavior data and text data at the same time. Using LUBD-CM, we are able to capture hidden behavior dimensions of users at a much finer granularity, which are especially effective in capturing their habits and characters. Then the credit profiling task can be done using standard $l2$-regularized logistic regression classifier whose input is the latent user behavior dimensions. After the

[2] https://www.kabbage.com/.
[3] http://www.zestfinance.com.

classifier learning phase, we can distinguish behavior dimensions that are informative for credit prediction from the classifier easily. By comparing with several state-of-the-art algorithms, we show that LUBD-CM has a much better predictive performance in terms of averaged accuracy, precision, recall and F1-Score, the most common measures in credit scoring. Besides, case studies show that the learnt latent behavior dimensions have excellent abilities in explaining users' credit label, which is very necessary in the practice of credit profiling.

The main contributions of this paper are as follows:

- Our work aims to infer the especially subtle and subjective user attribute – credit. We find that some types of user behaviors on social media are very informative for credit evaluation. To the best of our knowledge, we are the first to formally investigate personal credit profiling problem under the social media data setting.
- We propose a latent variable model LUBD-CM to incorporate as many as 5 different types of user behaviors with text data. In this way, we are able to capture latent user behavior dimensions from the social data at a much finer granularity.
- We conduct comprehensive experiments on a dataset crawled from Sina Weibo[4]. Experimental results demonstrate that LUBD-CM outperforms several state-of-the-art baselines and the learnt latent behavior dimensions are very interpretable for personal credit profiling.

The rest of the paper is organized as follows. In Sect. 2, we introduce the preliminaries and definition of personal credit profiling problem. In Sect. 3, we discuss our approach's framework and present the LUBD-CM model in detail. We present experimental results in Sect. 4. Finally, we review the related work in Sect. 5 and conclude the paper in Sect. 6.

2 Preliminaries and Problem Definition

Dataset Collection and Description. On twitter-style websites like Sina Weibo, one's online tweets are publicly available by nature. Generally speaking, anyone can access others' tweet data even if she is not a friend of the given user. Sina Weibo allows us to access and store one's all tweet data after we are granted with privileges by the given user. With the help of an online P2P lending partner, we obtain more than 200,000 users' Sina Weibo data, whose credit labels are known through the partner's internal data. These users have received at least one credit loan from the P2P lending company. Usually, the credit label is defined by whether the user has defaulted on any loan or not. That is, if the user defaulted on any loan transactions before, he or she is labeled as "bad credit"; if the user has never defaulted, he or she is labeled as "good credit". All users in the Sina Weibo testbed have authorized the company to collect their tweet data, which is a common prerequisite to make loans from the P2P lending companies. As a result, there are no privacy breaches or moral issues to study these users' credit risk based on Sina Weibo data (Table. 1).

[4] http://www.weibo.com, the most famous tweet-style platform in China.

Table 1. Some statistics of Sina Weibo Dataset for Credit Profiling.

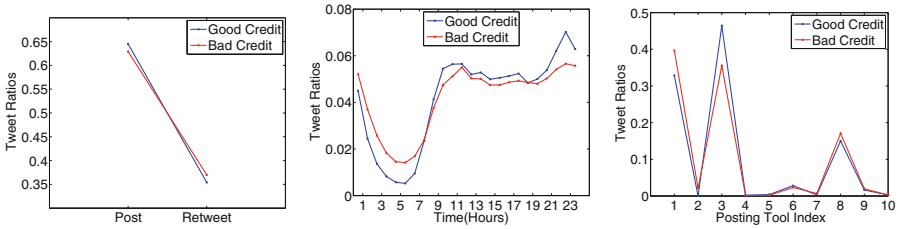
(a) Some Statistics of the dataset

Description	Value
# of good credit users	3,000
# of bad credit users	3,000
Total size of tweets	904,013
Total number of words	12,301,485
Size of vocabularies	241,197

(b) Summary of behavior types

Behavior Types	# of Possible Values
Retweet or not	2
Posting time(Hours)	24
Posting time(Days)	7
Posting tools	4012
# of emoticons	65

Adequate tweet data is crucial for algorithms' performance, so we set the minimum number of tweets for each user to be 10. Only users with no less than 10 tweets are chosen as experiment data. After removing users with less than 10 tweets, only 3,119 bad credit users are left. Therefore we randomly sample 3,000 good and 3,000 bad credit users from the filtered dataset to construct a balanced dataset for measuring the overall performance of LUBD-CM. For vocabularies, we remove stop words and infrequent words whose document frequency is less than 5. In Table 2(a), we summarize the main statistics of Sina Weibo Dataset. We consider as many as 5 different behavior types, including (1)*whether the tweet is retweeted or not,* (2) *hours of the day when the tweet is posted*, (3) *days of the week when the tweet is posted*, (4) *type of tools used to post the tweet*, (5)*number of emoticons*[5] *in the tweet*. In regards to the 5 behavior types, the number of their possible values ranges from 2 to as large as 4012, and details are listed in Table 2(b). Mostly, the distribution of each behavior type is different from each other, but all somehow follow the power-law distribution.



(a) Retweet or not (b) Posting Time (c) Posting Tool

Fig. 1. User distribution comparison between good and bad credit users w.r.t. behavior type "retweet or not", "posting time(Hours)" and "posting tools"

Motivating Examples. We demonstrate the motivation of exploiting behavior data by comparing the distribution differences between good and bad credit

[5] Icons expressing users' tempers and emotions.

users. Figure 1 examines the distribution difference between good and bad credit users with respect to three behavior types mentioned above. In Fig. 1(a), we can see that good credit users tend to post rather than retweeting compared to bad credit users. Good credit users are more likely to be a creator on the social media to some extent. In Fig. 1(b), there is a clear difference between good and bad credit users in terms of fraction of tweets posted at different hours of the day. Overall, people usually tweet between 9:00 and 24:00. But good credit users show tendency to post more during the daytime than bad credit users, while bad credit users are more likely to tweet during late night, which is an unhealthy lifestyle. It is reasonable that this behavior characteristic of bad credit users increases their risks to have medical emergencies, which may cause them to miss the payments. For the behavior type named "posting tools", we sample 10 representative posting tools. In Fig. 1(c), we can observe obvious differences for these posting tools, indicating that posting tool differences exist between good and bad credit users.

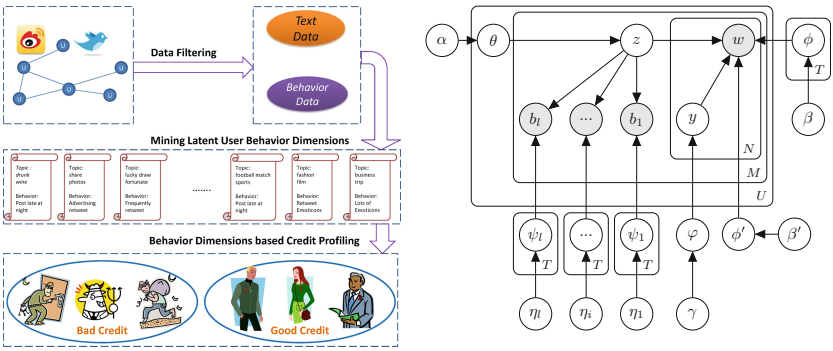
For the above mentioned behavior types, the differences between good and bad credit users all pass significance test at confidence level of 95 %. Similar results can be found with other behavior types. All the above observations validate that behavior data is informative and discriminative for credit prediction. Although the differences between good and bad credit users are very small, a combination of them can lead to a better result. It is worth mentioning that for many other behavior types like "time intervals between posts" or "usage of punctuation", there is no difference between good and bad credit users. Details of these behavior types are omitted due to space limitations.

Problem Definition. The definition of the problem we study can be formalized as follows: *Given a social data composed by $U \times N$ tweets that are generated by U users, our problem is to learn latent user dimensions $\Theta = \{\theta_u\}_{u=1}^{U}$ that can model users' tweet data at a high level, and infer the subjective attribute of each user's credit using these latent dimensions as features.* In order to achieve a considerable performance, we propose to take both text data and behavior data into consideration.

All notations used above can be found in Table 2. We note that almost all the literatures in credit scoring only formulate the credit scoring problem as a binary classification problem. Thus, only predicting whether a user's credit risk class is "good" or "bad" is enough for credit scoring. In addition, it is both impractical and inconvincible to directly assign credit scores to training samples. Usually, the credit score can be obtain after post-processing on the output of binary classifiers. We follow this convention in the study. Although our work aims to separate noises from tweet data for accurate social-data-based credit profiling, we acknowledge that it is very hard to predict users' default risk with a very high accuracy. We view the approach described here as a compliment to existing credit scoring methods. For instance, the latent dimensions we extracted can serve as auxiliary variables when financial data or survey data is also available. And we believe that results will become better when more social data are available from different social media websites.

3 Our Approach

In this section, we first introduce the framework of our approach for social-data-based credit profiling. Figure 2(a) shows our approach's framework, which first takes both behavior and text data into consideration for learning latent user behavior dimensions from social data, and then infer the credit risk label using standard classification algorithms. During the classification phase, all the learnt user behavior dimensions are treated as features. As illustrated in Fig. 2(a), the same behavior habit of posting at late night has different meanings when associated with topics of being drunk and watching football match respectively. Harnessing the behavior patterns inferred from both texts and behaviors of UGC, these latent behavior dimensions can predict whether a user is of good credit or bad credit more effectively. In the following, we will describe our LUBD-CM model that implements the framework of our credit profiling approach, which is also the core component of the framework.



(a) Framework of our approach (b) Plate notation for LUBD-CM model

Fig. 2. Framework of our credit profiling approach and plate notation for the LUBD-CM model

3.1 The LUBD-CM Model

To learn latent user behavior dimensions from both behavior and text part of social data, we propose a novel multiple behaviors enhanced topic model that integrates users' multiple behaviors simultaneously with textual content of tweets, called LUBD-CM (Latent User Behavior Dimension based Credit Model) for credit prediction. Figure 2(b) shows the plate notation for our proposed LUBD-CM model. The notations and their meanings are summarized in Table 2.

Modeling short text. The tweets, a kind of short text, are informal and heterogeneous. In [11], Hong et al. treat a given user's all tweets as a single pseudo

Table 2. Summary of notations and their meanings presented in Fig. 2(b)

Notations	Notation meanings	Notations	Notation meanings
U	# of users	ϕ	Topic word distribution
M	# of tweets of a user	$\psi_1, ..., \psi_l$	Topic behavior distribution of l types
N	# of words in a tweet	ϕ'	Background word distribution
T	# of topics	θ	User topic distribution
b_1,..., b_l	Behaviors of l types in a tweet	φ	Bernoulli distribution generating y
w	Word in a given tweet	$\alpha, \beta, \beta', \gamma$	Dirichlet priors
z	Topic of a given tweet	$\eta_1, ..., \eta_l$	Dirichlet priors for topic behavior distributions $\psi_1, ..., \psi_l$
y	Switch variable deciding whether or not to sample from ϕ'		

document and assume that words in the document are generated from a mixture of topics as LDA [2]. However, their study shows that traditional LDA topic features on tweets are not superior to TF-IDF features in twitter user classification. Following the ideas presented in [26], we assume that each tweet is generated from a single topic. And a tweet may contain both topic specific and background words[6] to handle the informality of tweets. Thereby, each word is generated with a switch variable y to determine whether it is generated by a background multinomial distribution or by a topic word multinomial distribution. Specifically, y follows the Bernoulli distribution. As the model in [26] is designed for tweet data analysis, it is also called TweetLDA. Neglecting the variables of multiple behaviors $b_1, b_2, ..., b_l$, LUBD-CM can be reduced to TweetLDA.

Modeling behavior data. Now we present the techniques used for modeling the behavior data. As behavior data is associated with each tweet, we assume that each behavior is generated after the topic variable z of the tweet is sampled. Each behavior is then sampled from a bag-of-behaviors distribution of the corresponding behavior type, which is also a multinomial distribution. In LUBD-CM, we assume that each tweet has multiple behaviors attached to it, as illustrated in Fig. 2(b). A similar behavior topic model for tweets is proposed by Qiu et al. [20], called B-LDA. Our LUBD-CM model is superior to B-LDA model in that (1) LUBD-CM is able to handle different types of user behaviors simultaneously; (2)With multiple behaviors, LUBD-CM obtains latent behavior dimensions to represent each user at a much finer granularity, which is very crucial for subtle attribute inference, like credit profiling.

[6] Background words are like stop words in tweets.

For model inference, we use the most widely adopted collapsed Gibbs sampling method [8] to infer the parameters of LUBD-CM model. Due to space limit, we omit the details of model inference and parameter estimation. Given the presented LUBD-CM model in Fig. 2(b), the generative process for both text and behavior data can be summarized as follows:

Algorithm 1. Generative Process for LUBD-CM

for *each topic* $t = 1, ..., T$ **do**
  Sample $\phi_t \sim Dir(\beta)$;
  Sample $\psi_{1,t} \sim Dir(\eta_1)$, ..., Sample $\psi_{l,t} \sim Dir(\eta_l)$;
Sample $\phi' \sim Dir(\beta')$;
Sample $\varphi \sim Dir(\gamma)$;
for *each user* $u = 1, ..., U$ **do**
  Sample topic distribution $\theta_u \sim Dir(\alpha)$;
  for *each tweet* $m = 1, ..., M_u$ *in user* u*'s all* M_u *tweets* **do**
    Sample a topic $z_{u,m}$ from θ_u;
    for *each word* $n = 1, ..., N_{u,m}$ **do**
      Sample $y_{u,m,n}$ from Bernoulli(φ);
      Sample $w_{u,m,n} \sim \phi'$ if $y_{u,m,n} = 0$, otherwise sample $w_{u,m,n} \sim \phi_{z_{u,m}}$;
    for *each behavior of type* l *associated with tweet* m **do**
      sample the behavior $b_{u,m_l} \sim \psi_{z_{u,m_l}}$;

4 Experiments

4.1 Experiment Setup

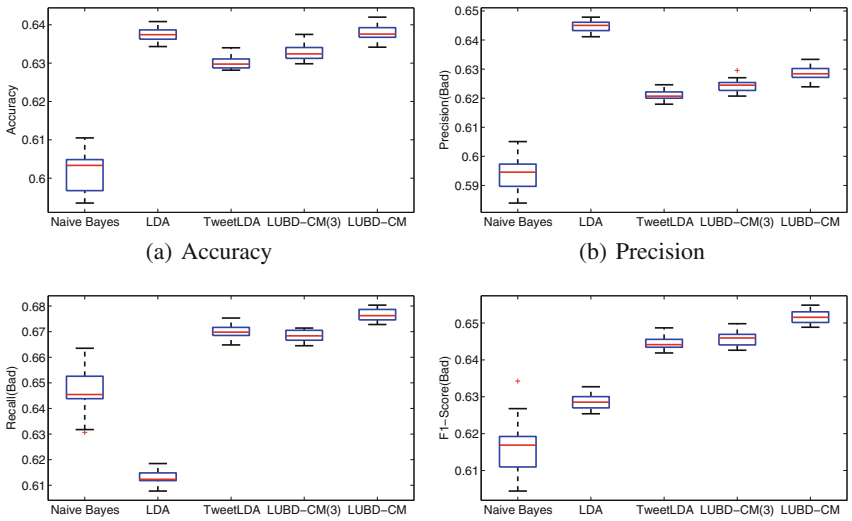
To compare LUBD-CM and the baselines' performance, we run 20 rounds of 10-fold cross validation. The classifier for credit prediction is l_2-regularized logistic regression [7], whose performance is the best in practice. Evaluation metrics include averaged accuracy, precision, recall and F1-Score. Since we are mostly interested in identifying the bad credit ones, the measures of precision, recall and F1-Score are computed based on the bad credit label.

We implement baseline methods including Naive Bayes, LDA, TweetLDA, and LUBD-CM(3). Specifically, Naive Bayes method corresponds to the traditional unigram features based method, which is very effective for user attribute classification [3,21], and LUBD-CM(3) is a variant of LUBD-CM that takes the first 3 types of behavior into account. Similar results can be observed for other cases of combining 3 behavior types. For Naive Bayes methods, no parameters are needed. For LDA, TweetLDA, LUBD-CM(3), and LUBD-CM, we find the optimal values for parameters T, β, β', γ and η using grid search with cross validation. During experiments, α is set to be $50/T$ where T is 150. Both β and β' are set to be 0.01, and γ is set to be 40. In LDA, β is set to 0.1, and in LUBD-CM, $\eta_i = \{0.01, 0.1, 0.1, 1, 1\}$ for $i = \{1, 2, 3, 4, 5\}$.

4.2 Experiment Results

Credit Prediction. Figure 3 shows performance comparison between Naive Bayes, LDA, TweetLDA, LUBD-CM(3), and LUBD-CM w.r.t. averaged accuracy, precision, recall, F1-Score respectively. From Fig. 3, we can clearly see that LUBD-CM consistently outperforms the baselines in credit prediction. We performed a t-test on different metrics, and showed that all the differences between LUBD-CM and baselines were statistically significant at confidence level of 95 %. This observation validates that it is superior to consider multiple types of behavior data for inferring user credit. Although the performance improvement of LUBD-CM over baselines is only about 1 %∼4 %, this performance improvement can contribute a multitude of revenues to P2P-lending companies in real life. It is worth noting that LDA has comparable performance with LUBD-CM in terms of Accuracy, but its F1-Score value is quite worse. Beside, the precision and recall values of LDA is quite different from other methods. The probable reason may lie in that LDA is usually not suitable for short-text like tweets.

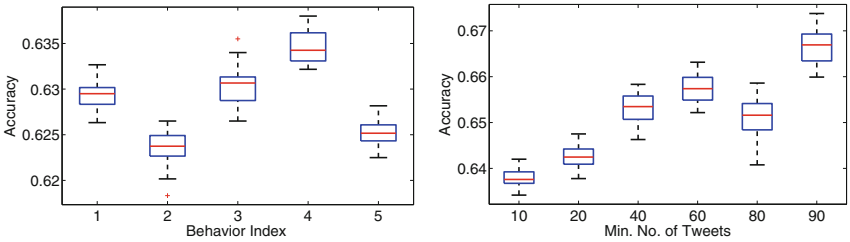
In Fig. 4(a), we show the performance comparison between 5 different LUBD-CM(1)s, which take only one behavior type into consideration. The results show that different behaviors have different impacts on credit prediction and only considering one behavior is not enough for credit profiling. Figure 4(b) shows the performance changes of LUBD-CM as the minimum number of tweets for each user increases. The overall increasing trend demonstrates that the more the data, the better the performance. And we can expect that if more social data per user is available, the performance of LUBD-CM will become even better.



(a) Accuracy (b) Precision

(c) Recall (d) F1-Score

Fig. 3. Performance comparison between LUBD-CM and baselines w.r.t. accuracy, precision, recall and F1-Score.



(a) Comparison of LUBD-CM(1)s (b) Min. No. of tweets

Fig. 4. Performance comparison between LUBD-CM(1)s and LUBD-CM's sensitivity to minimum number of tweets per user.

Case Studies of Latent Behavior Dimensions. After the classifier learning step, each feature, i.e., the latent user behavior dimension, is output with a weight indicating its predictive coefficient within the classifier. Utilizing these weights, we can identify the most predictive behavior dimensions. We find that dimension 29, 51, 6, and 90 are the four most important ones according to the weights associated with them. Among them, dimension 29 is negatively weighted, indicating its contribution to bad credit label, while the rest are positively weighted, indicating their contribution to the good credit label. We analyze the four latent user dimensions in detail as follows:

1. Dimension 29 includes words like "lucky draw", "prize", "money", and "ipad" etc. The probability of this dimension for retweeting and posting is 0.98 and 0.02 respectively, indicating that users of this dimension mostly retweet instead of posting. Users of this dimension often tweet late at night, at time between 3:00 AM∼4:00 AM. All these characteristics show that this behavior dimension is about retweeting advertising posts and wining prizes from lucky draws offered by the advertisers. We can infer that users of this dimension are not economically well off, desire for small bonuses and are more likely to miss the payments.
2. Dimension 51 contains words like "highway", "traffic", "jam" etc. The behavior distribution of "posting time" shows that users of this dimension often send tweets between 8:00 AM∼9:00 AM, indicating that they are on their way to work and a traffic jam happens. This kind of users often have stable employments, and their credit labels are therefore likely to be good.
3. Dimension 6 includes words like "nation", "society", "government" etc., indicating that users of this dimension often care about affairs related to society and government. From "posting time" behavior, we find that users of this type seldom stay up late in the night sending tweets, and other types of behaviors are all quite normal, indicating that they are ordinary people caring about the public affairs. With no anomaly behavior patterns and paying attention to public affairs, users of this dimension are usually responsible adults and more likely to have good credit.

4. Dimension 90's representative words are "enjoy", "film", "feeling", "tears" etc. The "posting time" behavior distributions on this dimension indicate that tweets of this dimension are often sent between 7:00 PM∼11:00 PM on Friday, Saturday or Sunday. This phenomena clearly shows that this dimension is about watching films in cinemas. Users of this kind are usually fond of spiritual consumption. And they are somehow intellectually well developed and seldom ruin their credit.

We also observe that for two dimensions (25, 76) composed by emoticons, their number of emoticons is mostly between 1 and 8 rather than 0. One of them represents the happiness emotions of users and contributes to the good credit label, while the other dimension indicates that users of this kind is very upset and sad and contributes to the bad credit label. This observation also coincides with human intuitions that good credit users shall be more optimistic and happier than bad credit users in real life.

5 Related Work

Social-data-based credit scoring can be viewed as inferring the specific user attribute named credit from social data, which is closely related to user profiling on social media. Rao et al. [21] firstly attempt to classify user attributes including gender, age, region and political affiliation based on features from tweets like unigram and bigram word features and sociolinguistic features. Pennacchiotti and Popescu [19] conducted study for user profiling on twitter with respect to political affiliation, ethnicity, and affinity to a certain brand with more diverse features. Other studies inferring users' attributes including gender [3], age [18], occupation [25] etc., also take advantage of tweet content. Besides, social connections between online users are also explored for user attribute inference in [5,17]. Taking one step further, Li et al. [16] proposed a user co-profiling methodology to model relationship types and user attributes simultaneously. Nonetheless, only text or network data are heavily leveraged in previous user profiling studies. Behavior data on the social web is neglected in most cases, though user behavioral patterns and habits could be very informative for user attribute profiling.

Our work is also related to traditional consumer credit scoring, which also focuses on small loans applied by individual consumers. Abundant research has been devoted to it based on statistical methods [4,9], including discriminant analysis [6], logistic regression [23], decision tree [1], neural networks [13] etc. Recent years have also witnessed the fast development of advanced methods for credit scoring [14,24]. In particular, Harris [10] assesses credit risks using optimal default definition selection algorithm, which selects the best default definition for building models. However, nearly all these works are based on transactional loan/payment records, credit reports or demographic survey data, which is crucially different from social-data-based personal credit scoring.

6 Conclusion

In this paper, we are purposed to harness the social data for personal credit profiling. We found that users' some kinds of behavior data benefits the task greatly, which also coincides with human intuitions. We proposed a joint topic-behavior model LUBD-CM to learn fine-grained latent user behavior dimensions. We conducted extensive experiments on a Sina Weibo dataset. Experimental results validated that our approach using latent dimensions inferred from LUBD-CM outperforms several state-of-the-art baselines with a significant margin. In the future, we plan to investigate more informative behavior types to boost LUBD-CM's performance. In addition, we'd like to improve our model's scalability to make it suitable for dealing with large-scale social data.

Acknowledgement. This research was partially supported by grants from the National Science Foundation for Distinguished Young Scholars of China (Grant No. 61325010), the National High Technology Research and Development Program of China (Grant No. 2014AA015203), the Science and Technology Program for Public Wellbeing (Grant No. 2013GS340302) and the CCF-Tencent Open Research Fund. This work was also partially supported by the Pinnacle Lab for Analytics @ Singapore Management University.

References

1. Arminger, G., Enache, D., Bonne, T.: Analyzing credit risk data: a comparison of logistic discrimination, classification tree analysis, and feedforward networks. Comput. Stat. **12**(2), 293–310 (1997)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
3. Burger, J.D., Henderson, J.C., Kim, G., Zarrella, G.: Discriminating gender on twitter. In: EMNLP, pp. 1301–1309 (2011)
4. Crook, J.N., Edelman, D.B., Thomas, L.C.: Recent developments in consumer credit risk assessment. Eur. J. Oper. Res. **183**(3), 1447–1465 (2007)
5. Dong, Y., Yang, Y., Tang, J., Yang, Y., Chawla, N.V.: Inferring user demographics and social strategies in mobile social networks. In: KDD, pp. 15–24 (2014)
6. Eisenbeis, R.A.: Problems in applying discriminant analysis in credit scoring models. J. Bank. Finance **2**(3), 205–219 (1978)
7. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J.: Liblinear: a library for large linear classification. J. Mach. Learn. Res. **9**, 1871–1874 (2008)
8. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. Natl. Acad. Sci. **101**(suppl 1), 5228–5235 (2004)
9. Hand, D.J., Henley, W.E.: Statistical classification methods in consumer credit scoring: a review. J. Royal Stat. Soc. Ser. A (Stat. Soc.) **160**(3), 523–541 (1997)
10. Harris, T.: Default definition selection for credit scoring. Artif. Intell. Res. **2**(4), 49 (2013)
11. Hong, L., Davison, B.D.: Empirical study of topic modeling in twitter. In: Proceedings of the First Workshop on Social Media Analytics, pp. 80–88. ACM (2010)

12. Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: understanding microblogging usage and communities. In: WebKDD/SNA-KDD, WebKDD/SNA-KDD 2007, pp. 56–65 (2007)
13. Jensen, H.L.: Using neural networks for credit scoring. Manag. Finance **18**(6), 15–26 (1992)
14. Kruppa, J., Schwarz, A., Arminger, G., Ziegler, A.: Consumer credit risk: individual probability estimates using machine learning. Expert Syst. Appl. **40**(13), 5125–5131 (2013)
15. Kwak, H., Lee, C., Park, H., Moon, S.: What is twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World Wide Web, WWW, pp. 591–600 (2010)
16. Li, R., Wang, C., Chang, K.C.-C.: User profiling in an ego network: co-profiling attributes and relationships. In: Proceedings of the 23rd International Conference on World Wide Web, WWW (2014)
17. Mislove, A., Viswanath, B., Gummadi, P.K., Druschel, P.: You are who you know: inferring user profiles in online social networks. In: WSDM, pp. 251–260 (2010)
18. Nguyen, D., Gravel, R., Trieschnigg, D., Meder, T.: "How old do you think i am?" a study of language and age in twitter. In: ICWSM (2013)
19. Pennacchiotti, M., Popescu, A.-M.: Democrats, republicans and starbucks afficionados: user classification in twitter. In: KDD, pp. 430–438 (2011)
20. Qiu, M., Zhu, F., Jiang, J.: It is not just what we say, but how we say them: Lda-based behavior-topic model. In: SDM, pp. 794–802 (2013)
21. Rao, D., Yarowsky, D., Shreevats, A., Gupta, M.: Classifying latent user attributes in twitter. In: SMUC, pp. 37–44 (2010)
22. Rosenthal, S., McKeown, K.: Age prediction in blogs: a study of style, content, and online behavior in pre- and post-social media generations. In: ACL, pp. 763–772 (2011)
23. Wiginton, J.C.: A note on the comparison of logit and discriminant models of consumer credit behavior. J. Financial Quant. Anal. **15**(03), 757–770 (1980)
24. Yap, B.W., Ong, S.H., Husain, N.H.M.: Using data mining to improve assessment of credit worthiness via credit scoring models. Expert Syst. Appl. **38**(10), 13274–13283 (2011)
25. Zeng, G., Luo, P., Chen, E., Wang, M.: From social user activities to people affiliation. In: ICDM (2013)
26. Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., Li, X.: Comparing twitter and traditional media using topic models. In: ECIR, pp. 338–349 (2011)