Weibo user attribute analysis method based on multi-feature

KeFei Cheng¹, Fengchi Shan*, ZhiWen Song¹, ZhenNi Zhao¹, XiaoYong Guo¹

¹ College of Computer Science & Technology, Chongqing University of Posts and Telecommunications Chongqing 400065, China

*Corresponding author’s e-mail: shanfengchi@qq.com

Abstract. In view of the deficiency of domestic Weibo user attribute analysis, the imperfection of Weibo feature extraction and the problem that the classification accuracy needs to be improved, a Weibo user attribute analysis method based on multi-features is proposed. This paper first uses the word2vec model to build text features from Weibo content, then constructs Weibo user features from Weibo information and user information, and finally sends multi-feature sets into the improved three-tier stacking model to build Weibo user attribute analysis model. The experimental results show that this method has better classification effect than the text-based classification method and the traditional stacking model.

1. Introduction

With the increasing popularity of online social media, online information has become huge and complex. With the help of computer technology, it has become an important topic for both academia and industry to deeply understand the basic information of individuals and groups, explore social psychology and behavior patterns, quickly and accurately provide personalized and multifaceted decision support, and assist in solving practical social problems. The deep understanding of user information and user behavior is one of the core contents. As personal attribute data often involves privacy issues, users often choose not to fill in or fill in false information to hide their personal information, resulting in the basic information related to users usually cannot be obtained directly. User attribute analysis can solve this kind of problem.

At present, the study on user attributes of social media platforms is mainly carried out on Twitter, Facebook, Instagram and other platforms. José Ahirton Batista Lopes Filho et al.[1] extracted 60 features from Twitter text and used three machine learning algorithms to identify the gender of users; Take Yo et al.[2] constructed word vectors through word2vec model, and used machine learning algorithms to identify users’ personal attributes. Simaki et al.[3] mined the text information in Twitter and obtained a set of text features based on sociolinguistics, text mining and content relevance, and inferred the user’s age by using random forest algorithm.

The main researches on the attribute analysis of Chinese Weibo users are as follows: BaoGin Liu et al.[4] constructed a SVM classifier by using the features of emotional words and the language style features related to emotion, thus realizing the gender classification of Weibo users. Pu Zhang et al.[5] firstly extracted a series of artificial features from Weibo text data to build a classifier, and then automatically extracted features from the convolution neural network model to build a classifier. Finally, the user gender classifier was obtained by using the XGBoost model to fuse the two classifiers. Qiu Jing et al.[6] combined with the characteristics of Chinese Weibo text data, reprocessed the data based on Weibo user granularity, and then used Logistic Regression, Random Forest, SVM and other models for classification. Finally, the method of fusion of multiple SVM models was used to improve the gender classification accuracy of Weibo users.
To sum up, because the information used in the research of user attribute analysis is too single, there is no comprehensive use of a variety of information features, and the model used is more traditional, so the traditional methods will no longer be suitable for Weibo.

In view of the shortcomings of traditional user attribute analysis methods, this paper takes Sina Weibo as the research object and proposes a user attribute analysis method based on multi-feature fusion. Firstly, the text features are constructed from Weibo text data by word2vec model; secondly, the basic features for Weibo attribute analysis are constructed from Weibo user data, and the compound features that accord with Weibo users are constructed through the basic features. Finally, the Weibo user attribute analysis model is constructed by using the improved three-layer Stacking algorithm. Experimental results show that this method can effectively improve the classification effect of Weibo user attributes.

2. Related Knowledge

2.1. Word2vec model

Word2vec[7] is a word vector generation model based on neural network published by google researcher. The model uses deep learning network to model the semantic relationship between words and their contexts of corpus data in order to obtain dimensional word vectors. The vector is generally around 100-300 dimensions, which can well solve the problem of high-dimensional sparsity of the traditional word vector space model. Because the deep neural network model can abstract the high-level semantics of features, the model can avoid the semantic gap. Therefore, Word2vec is an excellent algorithm in natural language processing at present.

There are two training models in Word2vec, which are CBOW (Continuous Bag-of-Words) model and Skip-gram model. As shown in figure 1, the CBOW model changes the known words at the input layer \( w(t - m), \ldots, w(t - 1), w(t + 1), \ldots, w(t + m) \), It is converted to a word vector, and then the projection layer sums the word vectors in the input layer, that is, the vectors are added. Then the output layer uses SoftMax regression[8] to output the probability that the target word is a word \( P(w_t | w_{t-m}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+m}) \). The word2vec model models the corpus, that is, to determine the word vector \( w_t = (w_{v_1}, w_{v_2}, w_{v_3}, \ldots, w_{v_m}) \) corresponding to each word as the prediction feature, so that the CBOW model predicts the unknown words by knowing the context of the word, while the Skip-gram model predicts the unknown words in the context through the known words.

![Figure 1. CBOW and Skip-gram model](image-url)
2.2. Stacking algorithm
Suppose there are N learning methods on the data set D, which is composed of training sets. The N learning algorithms get N classifiers and become base classifiers. Each sample in the data set D will output N results after the N classifiers. The N results are taken as new samples, and the sample label corresponds to the sample label in the data set D. The new sample is used to train a classifier. The classifier becomes a meta-classifier. Figure 2 shows the structural framework of the Stacking algorithm. When using Stacking algorithm[9], the data is classified by N base classifiers, and the N results are formed into new vectors and then classified by meta-classifier to get the final result.

![Figure 2. Stacking model framework](image)

3. User attribute inference method based on Multi-feature Fusion

3.1. Text feature construction
User blog posts record the daily life of users and can effectively depict the attributes and characteristics of users. Therefore, we use the Jieba word segmentation tool to segment the sample, remove the stopped words, and merge each user's Weibo, and get a collection of user blog posts $m_i = \{T_{m_1}, T_{m_2}, T_{m_3}, ..., T_{m_n}\}$, $m_i$ denotes the Weibo collection with user ID i, $T_{m_i}$ denotes a Weibo collection of individual users, $T_{m_i} = \{w_1, w_2, w_3, ..., w_t\}$, $w_t$ denotes a word for a single Weibo.

Weibo users' Weibo is trained by Skip-Gram model to obtain 300-dimensional word vectors in Weibo. On this basis, the Weibo vector of each user is calculated, and the formula is as follows.

$$\text{userVec}(u_i) = \frac{\sum_{k=1}^{K} Wvec_k}{K}$$

Where $u_i$ denotes the user whose ID is i, $K$ denotes the number of Weibo words of user $u_i$, and $Wvec_k$ denotes the word vector of the k th word.

The Stacking model is used as the combination strategy of integrated learning, and the support vector machine (SVM), decision tree (decision tree), logical regression (Logistic), optical gradient hoist[10](LightGBM) and extreme gradient lifting[11](XGBoost) are used as primary classifiers. The prediction results are obtained by the combination of logical regression (Logistic) as a two-layer classifier, and the text feature classification results are obtained. The classification results are shown in Table 1.

| Gender analysis | Age analysis |
|-----------------|--------------|
| Precision/%     | Recall/%     | F1/%         | Precision/% | Recall/% | F1/% |
| 75.56           | 70.23        | 73.23        | 74.10       | 72.39    | 73.79 |

From Table 1, we can see that only relying on Weibo text to build a user attribute classifier cannot achieve a better classification effect, so the following will combine user characteristics and improve the traditional two-layer Stacking model to improve the effect of user attribute classifier.
3.2. User feature construction

3.2.1. Common feature construction

Starting from the dimensions of user blog post, user profile and user fan association table, this paper uses a variety of feature types to analyze user attributes, and the common features are shown in Table 2.

Table 2. Common features

| Feature category         | Specific features                                                                 |
|-------------------------|-----------------------------------------------------------------------------------|
| Time characteristics    | The number of Weibo in each time period, The number of Weibo on each of the seven days of the week |
| Numerical characteristics| Maximum number of retweets, minimum number of retweets, average number of retweets, median number of retweets, maximum number of comments, minimum number of comments, mean number of comments, median comments, Weibo comment rate, average number of Weibo words, minimum number of Weibo words |
| Statistical characteristics| Number of Weibo, number of followers, number of followers, number of cross-relations, number of comments, number of retweets |
| Content features        | Weibo length, photo, URL, username length, registration location, profile, education information, nickname, avatar |

3.2.2. Compound feature construction

3.2.2.1. User activity

The calculation formula of user activity feature $f_{\text{useractive}}(u_i)$ is as follows:

$$f_{\text{useractive}}(u_i) = \frac{f_{\text{sum}}(u_i) - f_{\text{transpond}}(u_i)}{f_{\text{time}}(u_i)} \quad (2)$$

Where $u_i$ denotes the user of $i$, $f_{\text{sum}}(u_i)$ denotes the total number of Weibo of $u_i$, $f_{\text{transpond}}(u_i)$ denotes the number of Weibo retweets of user $u_i$, and $f_{\text{time}}(u_i)$ denotes the time interval between the first Weibo and the last Weibo posted by user $u_i$.

3.2.2.2. Time distribution of user Weibo

The formula for calculating the time distribution $f_{\text{timedistribution}}(u_j)$ of users' Weibo is as follows:

$$f_{\text{timedistribution}}(u_j) = f_{\text{count}}(u_j) - f_{\text{transpond}}(u_j) \quad (3)$$

Where $u_j$ denotes the user whose ID is $i$ in the time period $j$, $f_{\text{count}}(u_j)$ denotes the number of Weibo posted by the user $i$ at the time $j$, and $f_{\text{transpond}}(u_j)$ denotes the number of Weibo forwarded by the user $i$ at the time $j$;

3.3. Three-layer Stacking classification model

Ensemble learning is a method in the field of machine learning, which can achieve better classification or regression effect by combining several weak supervision models into a strong supervision model. This section improves the traditional two-layer Stacking model, using different input methods for different features, so that the model can learn text features and user features more accurately, and select the boosting strong supervision model as the base classifier in the integration to improve the accuracy of user attribute prediction. The model structure is shown in figure 3:
In the method shown in Figure 3, firstly, the text features calculated in Section 3.1 are taken as the first input layer of the model, and the user attributes are preliminarily classified with support vector machine, decision tree, logical regression, optical gradient lifting tree and extreme gradient lifting tree respectively. Then, the initial classification results of the five classifiers are integrated, and the classification results are taken as features, and the second classification of user attributes is carried out by using logical regression. Finally, combining the secondary classification results with the user characteristics calculated in Section 3.2, we use the extreme gradient lifting tree for the final classification and get the final user attribute prediction results.

4. Experimental results and evaluation

4.1. Data set
This paper uses the data set released by SMP CUP 2016 as the experimental data, which is the real data set of Sina Weibo, including the social network composed of about 2.567 million Weibo users; about 46000 users' Weibo text, retweeted comments, time and other information; user attribute information describes the nicknames and avatars of about 2100 users.

There are 3200 tagged users in the official data, including two types of tags in the gender test task, including "f" and "m", and three types of tags in the age test task, including "before 1979", "1980-1989" and "1990". Use this part of the data as the training set, while the test set has 1240 untagged users. In this paper, based on the construction of user attribute feature system of 4440 users, as well as the training and prediction of user attribute analysis model, Weibo information in the data set is used to train Word2vec word vectors.

4.2. Evaluation index
Figures In this paper, we will measure it by accuracy P (Precision), recall rate R (Recall), and comprehensive consideration index F1.

\[
P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN}, \quad F1 = \frac{2PR}{P + R}
\]

In the above formula, \(TP\) represents the number of positive cases correctly classified, \(FN\) represents the number of positive cases mistakenly divided into negative cases, and \(FP\) represents the number of negative cases mistakenly divided into positive cases.
4.3. Result analysis

In order to verify the effectiveness of the Weibo user attribute analysis method based on multi-features proposed in this paper, the proposed method is compared with the following methods. The specific experiments are as follows:

Method 1: Use text features and use three-tier Stacking model to classify;
Method 2: The user characteristics are adopted and the three-layer Stacking model is used for classification;
Method 3: Text features and user features are adopted, and decision tree model is used for classification;
Method 4[6]: Text features and user features are adopted, and two-layer Stacking model is used for classification;
Method 5[12]: Text features are adopted and SVM model is used for classification;
Method 6[13]: Use text features to classify using a two-layer Stacking model;
In this paper: Text features and user features are adopted, and three-tier Stacking model is used for classification.

The experimental structure of the above seven different methods in gender and age analysis is shown in Table 3.

| Method       | Gender analysis | Age analysis |
|--------------|-----------------|--------------|
|              | Precision/%     | Recall/%     | F1/%         | Precision/% | Recall/% | F1/% |
| Method 1     | 76.39           | 72.84        | 73.57        | 76.21       | 76.31    | 74.53 |
| Method 2     | 83.61           | 86.25        | 82.39        | 81.22       | 79.85    | 80.12 |
| Method 3     | 62.56           | 62.80        | 62.63        | 84.35       | 84.50    | 84.41 |
| Method 4     | 75.88           | 86.81        | 80.98        | 80.52       | 67.63    | 51.26 |
| Method 5     | 66.92           | 73.63        | 70.13        | 65.55       | 74.15    | 72.23 |
| Method 6     | 75.56           | 70.23        | 73.23        | 74.10       | 72.39    | 73.79 |
| In this paper| 92.50           | 92.00        | 92.00        | 88.21       | 89.33    | 89.02 |

By comparing method 1, method 2 and this method, we can find that the model with user characteristics can effectively improve the effect of user attribute analysis. The use of user-based features can better reflect the characteristics of users, which is also in line with the characteristics of Weibo as a social network, but with the increase of user characteristics, it will increase the difficulty of feature engineering.

By comparing method 1, method 5 and method 6, we can find that the effect of using text feature as original feature and using three-layer Stacking model is better than traditional Stacking and SVM model. By comparing method 3, method 4 and this method, it can be found that this method can improve the accuracy and stability of user attribute classification when using text features and user features as original features and using three-layer Stacking model for training and user attribute classification.

Based on the analysis of the above results, it is concluded that the Weibo user attribute analysis method based on multi-features proposed in this paper is the best in feature comparison and method comparison experiments, indicating that this method is feasible in user attribute analysis.

5. Conclusion

In this paper, a Weibo user attribute analysis method based on multi-feature fusion is proposed. According to the uniqueness of Weibo text, the common features of Weibo users are constructed, and on this basis, the compound features suitable for Weibo users are constructed. This method also improves the traditional Stacking model to a three-tier model. Experimental results show that this method can effectively analyze the user attributes of Weibo. However, the disadvantage of this method is that it has high time complexity and is not suitable for real-time analysis system. Future work will focus on reducing the time complexity of the method and improving the execution efficiency of the method.
Acknowledgments
This work was supported in part by the National Natural Science Foundation of China (NO.17XFX013).

References
[1] Filho, J.A.B.L., R. Pasti and L.N.D. Castro, Gender Classification of Twitter Data Based on Textual Meta-Attributes Extraction. 2016.
[2] Yo, T. and K. Sasahara. Inference of personal attributes from tweets using machine learning. in 2017 IEEE International Conference on Big Data (Big Data). 2017: IEEE.
[3] Simaki, V., I. Mporas and V. Megalooikonomou. Age identification of twitter users: classification methods and sociolinguistic analysis. in International Conference on Intelligent Text Processing and Computational Linguistics. 2016: Springer.
[4] Liu, B. and Y. Niu, Gender recognition of chinese micro-blog users based on emotion features. Comput Eng Sci, 2016. 38(9): p. 1917-1923.
[5] Pu, Z., et al., Gender classification method for Chinese micro-blog users based on two classifier fusion. Computer Engineering and Design, 2019.
[6] Qiu, J., L.I. Lin and Y.P. Zheng, Research on the Gender Classification of Micro-blog Users. Journal of Liaoning University of Technology, 2018.
[7] Mikolov, T., et al. Distributed representations of words and phrases and their compositionality. in Advances in neural information processing systems. 2013.
[8] She, X. and Y. Zhu. Text Classification Research Based on Improved SoftMax Regression Algorithm. in 2018 11th International Symposium on Computational Intelligence and Design (ISCID). 2018: IEEE.
[9] Wolpert, D.H., Stacked Generalization. Neural Networks, 1992. 5(2): p. 241-259.
[10] Ke, G., et al. Lightgbm: A highly efficient gradient boosting decision tree. in Advances in Neural Information Processing Systems. 2017.
[11] Chen, T. and C. Guestrin. XGBoost: A Scalable Tree Boosting System. in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016.
[12] Zhang, D., et al., Chinese comments sentiment classification based on word2vec and SVMperf. Expert Systems with Applications, 2015. 42(4): p. 1857-1863.
[13] Penghui, M.A., et al., Research on User Portrait Construction. Modern Information Technology, 2019.