Multimodal Fusion for Image and Text Classification with Feature Selection and Dimension Reduction

Xinran Liu\textsuperscript{1,2,3}, Zhongju Wang\textsuperscript{1} and Long Wang\textsuperscript{1*}

\textsuperscript{1}School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China
\textsuperscript{2}Shunde Graduate School of University of Science and Technology Beijing, Shunde 528300, China
\textsuperscript{3}Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing 100083, China
\textsuperscript{*}Corresponding author’s e-mail: lwang@ustb.edu.cn

Abstract. Internet has become an important information platform, and it is very important to accurately understand the multimedia information of the Internet. In this paper, our main task is to do classification based on pictures and texts collected from the Internet, which is a classification problem of multimodal fusion in practice. However, when multimodal data is put together, there may occur the dimension disaster problem. We apply feature selection (FS) and dimension reduction (DR) in feature levels both in later fusion and early fusion to solve this problem. The classification accuracies in different models obtain improvements in different levels respectively. We also discuss the relation between single modals and multimodal in later fusion. In our experiments, images and text can be classified by multimodal models under FS/DR, and of which with the help the multimedia information from the Internet can be analysed better to help enterprises provide better services and products, and then carry out better network marketing and promotion.

1. Introduction
With the rapid popularization of mobile Internet [1], it was more convenient for users to use social network media. Hundreds of millions of users can share and exchange ideas with each other in time through various social network media, such as Weibo [2], WeChat [3], and Facebook [4], and participate in some hot events. In addition to traditional text information, current users can upload multimedia information such as pictures, sounds, and videos, which greatly enriches the content of social network media. If you can make full use of this content, you can better understand the interests of different users, design an accurate recommendation system, and improve the user experience. Besides, this user interest information is also of great significance to other fields. For example, the apparel industry can capture market trends based on current user interests [5], and the retail industry can understand market demand based on user interests, and adjust product types and supply quantities on time [6], the advertising industry can use user interests to target personalized advertisements. Therefore, using images and text information in social media to infer user interests has important research significance and application value.

With the increasing demand of social media users for the experience of using social media, personalized recommendation systems based on user interests have important use value for many social
media platforms, and therefore have received more research. Schenkel et al. [7] developed a personalized recommendation system using social relationships and an increasable threshold method. Nakamoto et al. [8] proposed a tag-based context-sensitive collaborative filtering method, which makes full use of user tag information. In addition to individual tag information, users’ social attributes are also often used as personalized recommendations. Qian et al. [9] and Yang et al. [10] use the method of tensor decomposition to establish a personalized user item recommendation system based on social tags. Cai et al. [12] explored the personalized search function through the user profile in the focus classification. In addition, the interaction between users and items is often used in personalized recommendation systems. Zhang et al. [13] constructed a three-part graph of user-item-label, and improved the use effect of the recommendation system through a recommendation algorithm based on diffusion. Iwata et al. [14] proposed a cross-domain recommendation method based on an unsupervised many-to-many object matching algorithm.

The above research puts forward effective solutions to the different problems in the personalized recommendation system. However, there is a lack of analysis of the content of pictures uploaded by users. In addition, the research on the recommendation method of social network pictures is still insufficient. The analysis and understanding of image content can effectively supplement the materials used in the personalized recommendation system and further improve the effect of personalized recommendation.

In this paper, we experiment on a dataset crawled from Pinterest. Pinterest is a photo-sharing social networking website where users could share and collect pictures. Users download, share and transmit pictures (aka pins) on this free website. Users can pin pins to their pinboards and classify pins themselves. However, not all pinboards are classified.

We use multimodal models to classify users and infer the interests of users. Considering that the fusion model may bring redundant information and the dimension reduction problem. We apply feature selection and dimension reduction techniques in the multimodal fusion model. The specific research content is as follows:

(1) Feature extraction is performed in two different ways, traditional feature extraction and neural network feature extraction. Support vector machine is employed to measure accuracy of different ways, and the experimental results of image classification and text classification are compared.

(2) For the problem of multimodal fusion, two different fusion methods are used. One is direct concatenation of image and text features, and the other is the weighted fusion of image and text prediction probability results for experiments.

(3) The interests of users are modelled based on category information combined with images and texts is used as the points of interest of users. The effectiveness of the proposed method is verified on the real social media dataset.

2. Materials and Methods

This paper developed multimodal fusion models for users and infer interest classification problem. Considering that fusion models may bring noise information and curse of dimensionality problems. Feature selection and dimension reduction techniques are applied in the multimodal fusion model. The specific research content is as follows:

(1) Feature extraction is performed in two different ways, traditional feature extraction and neural network feature extraction. Support vector machine is employed to measure accuracy of different ways, and the experimental results of image classification and text classification are compared.

(2) For the problem of multimodal fusion, two different fusion methods are used. One is direct concatenation of image and text features, and the other is the weighted fusion of image and text prediction probability results for experiments.

(3) The interests of users are modelled based on category information combined with images and texts is used as the points of interest of users. The effectiveness of the proposed method is verified on the real social media dataset.

2.1 Feature Representation

For images representation, it is obvious that CNN can perform the state-of-the-art results in image classification. Therefore, we complete the feature extraction on the basis of a CNN model known as Alexnet. In particular, we use the model trained on the ImageNet dataset for the image classification. Technically, we feed images to this feature extractor and chose the features of the Fc7 output as features of the image, which are 4096 dimensions.

Word Embedding is a most commonly used text representation model. In this paper, we use Embedding to extract text feature. The word vector model is a model that considers the positional
relationship of words. Through the training of a large number of corpora, each word is mapped into a high-dimensional vector. We use the vectors trained on three million words and phrases based on a Google News dataset. Each word will be transformed to vectors of 300 dimensions with these vectors.

2.2 Multimodal Fusion
Multimodality refers to a combination of two or more modalities in various forms. Each form of information is a modality. Since different modes have different expressions and different perspectives on things, there are some crossover and complementary phenomena, and there may even be a variety of different information interactions between modes. If multimodality can be handled reasonably, Information, it can get more feature information. Multimodal fusion is responsible for combining the information of multiple modalities for target prediction (classification or regression). The typical fusion methods are early fusion and later fusion.

Early fusion is to concatenate different features on feature-level and input them into model. Different from the early fusion, later fusion focuses on the individual learning of various modal data, and fusing the results of each modal learning into a multi-modal semantic representation. The later fusion fuses different modalities based on score-level. It trains multiple models, and each model has a prediction score. We fuse the results of all models to get the final prediction result. Common late fusion methods include average, maximum, and weighted average of scores, as well as logistics regression.

Although multimodal fusion improves the complementarity of information, it also brings redundant information and curse of dimensionality problems. Feature selection and dimensionality reduction of the fusion features are beneficial to reduce the complexity of the model and reduce the computational overhead. On the one hand, the feature selection methods select some features from all the extracted features as the training set features. The feature does not change the value before and after the selection, but the dimension of the feature after selection is definitely smaller than before selection. Algorithms relied on this strategy include binary particle swarm optimization (BPSO) [15], binary Jaya (BJaya) [16], Variance Feature Selection (VFS) [17], K Best Feature Selection (KFS), Fpr Feature Selection (FFS) [18], and Model Feature Selection (MFS). On the other hand, dimensionality reduction methods reduce the dimensionality of features by mapping features from high-dimensional space to a relatively low-dimensional, such as local linear embeddings (LLE), Principal Component Analysis (PCA) [19], Kernel PCA (KPCA) [20], and Latent Semantic Index (LSI) [21].

3. Experimental Results & Discussion
We crawled over 40,000 pictures from 32 classes according to the data list in [22]. We pick 5 classes according to number of pictures from all data to avoid the imbalance of data. In each class, we randomly select 1000 pictures and its corresponding texts to test performance of all considered models. Regarding the hyperparameters, we use five-fold cross validation on the training dataset and determine models. Classification accuracy of different models is listed in Table 1.

According to the results presented in Table 1, It's obvious that we have improved the accuracies of the classification using these techniques. Among the models of feature selection, BPSO has a good performance, and achieves an accuracy of 0.748 in the later fusion and an accuracy of 0.74 in early fusion. In the dimensionality reduction models, LLE can achieve an accuracy of 0.742 in later fusion and an accuracy of 0.744 in early fusion. As the accuracy of later fusion continues to increase, the accuracies of corresponding single modals all improve overall.

| Modal Algorithm | Single image | Single text | Early fusion | Later fusion |
|-----------------|--------------|-------------|--------------|--------------|
| Initial         | 0.6          | 0.632       | 0.708        | 0.718        |
| BPSO            | 0.63         | 0.658       | **0.74**     | **0.748**    |
| BJaya           | 0.638        | 0.66        | 0.742        | 0.73         |
Based on the results of previous experiments, we randomly select 20 images and corresponding texts from the test dataset. We use later fusion with feature selection based on PSO and use logistic regression to classify them. Distribution of user data under each class is shown in Figure 1. In Figure 1, the abscissa is enclosed in a circle, and the five discrete points represent five different categories. The ordinate represents the probability distribution of user data for each user under the current category.

| Method | VFS  | KFS  | MFS  | FFS  | LLE  | PCA  | KPCA | LSI  |
|--------|------|------|------|------|------|------|------|------|
|        | 0.604| 0.624| 0.616| 0.642| 0.618| 0.618| 0.616| 0.616|
|        | 0.642| 0.648| 0.638| 0.718| 0.662| 0.648| 0.652| 0.652|
|        |      |      |      |      |      |      |      |      |
|        | 0.718| 0.722| 0.718| 0.722| 0.744| 0.744| 0.726| 0.726|
|        |      |      |      |      |      |      |      |      |
|        | 0.726| 0.728| 0.728| 0.726| 0.742| 0.742| 0.728| 0.728|

![Figure 1. Interests distribution of Users.](image)

It’s obvious that user1 is more interested in “diy_crafts” and “film music books” while user2 prefer “design”. user3 likes “art” best and user4’s favorite category is “food drink”. After predicting the users’ interest, the website can make recommendations based on this information respectively. According to the distribution of these classification results, the user’s interest can be well predicted. At the same time, these classification results can also provide a good foundation for user recommendation based on recommendation algorithm. Better classification results frequently refer to more accurate predictions, which shows that our research has great practical significance.

4. Conclusions
In this article, we apply six methods of feature selection and four methods of dimensionality reduction to later fusion and early fusion of images and texts. These methods can improve the classification results of multimodal fusion more or less. We analyse the relationship between single-modal and multi-modal fusion classification results, and discuss the relationship between dimensions and accuracy. At the same time, we compare the accuracy of different models with the corresponding time. Finally, we predict the user interest based on the multimodal fusion classification with feature selection based on PSO.

In the future work, we will try to use better methods to extract features of images and texts, and apply feature selection and dimensionality reduction techniques to other early fusion methods. In addition, we will further explore the relationship between the results of single modals and multimodal in early fusion.
Acknowledgments
This work was supported by in part by the National Key R&D Program of China under Grant 2018YFC0810601, and in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2020A1515110431.

References
[1] A. Aijaz, M. Dohler, A. H. Aghvami, V. Friderikos, and M. Frodigh (2017). Realizing the Tactile Internet: Haptic Communications over Next Generation 5G Cellular Networks, IEEE Wirel. Commun., vol. 24, no. 2, pp. 82–89, Apr.
[2] C. H. Lien and Y. Cao (2014). Examining WeChat users’ motivations, trust, attitudes, and positive word-of-mouth: Evidence from China, Comput. Hum. Behav., vol. 41, no. Supplement C, pp. 104–111, Dec.
[3] S.-E. Kim, K. Y. Lee, S. I. Shin, and S.-B. Yang (2017). Effects of tourism information quality in social media on destination image formation: The case of Sina Weibo, Inf. Manage., vol. 54, no. 6, pp. 687–702, Sep.
[4] P. Luarn, H.-C. Kuo, H.-W. Lin, Y.-P. Chiu, and Y.-C. Jhan (2017). Analyzing User Preferences Using Facebook Fan Pages, Interfaces, Nov.
[5] Y. G. Cinar, S. Zoghbi, and M. F (2015). Moens, Inferring User Interests on Social Media from Text and Images, in 2015 IEEE International Conference on Data Mining Workshop (ICDMW), pp. 1342–1347.
[6] J. Kang and H. Lee (2017). Modeling user interest in social media using news media and wikipedia, Inf. Syst., vol. 65, no. Supplement C, pp. 52–64, Apr.
[7] R. Schenkel et al. (2008). Efficient Top-k Querying over Social-tagging Networks, in Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, pp. 523–530.
[8] R. Y. Nakamoto, S. Nakajima, J. Miyazaki, S. Uemura, H. Kato, and Y. Inagaki (2008). Reasonable Tag-based Collaborative Filtering for Social Tagging Systems, in Proceedings of the 2Nd ACM Workshop on Information Credibility on the Web, New York, NY, USA,, pp. 11–18.
[9] X. Qian, H. Feng, G. Zhao, and T. Mei, Personalized Recommendation Combining User Interest and Social Circle (2014). IEEE Trans. Knowl. Data Eng., vol. 26, no. 7, pp. 1763–1777, Jul.
[10] X. Yang, H. Steck, and Y. Liu (2012), Circle-based Recommendation in Online Social Networks, in Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, pp. 1267–1275.
[11] D. Rafaillidis, A. Axenopoulos, J. Etzold, S. Manolopoulos, and P. Daras (2014). Content-based Tag Propagation and Tensor Factorization for Personalized Item Recommendation Based on Social Tagging, ACM Trans Interact Intell Syst, vol. 3, no. 4, p. 26:1–26:27, Jan.
[12] Y. Cai, Q. Li, H. Xie, and H. Min, (2014). Exploring personalized searches using tag-based user profiles and resource profiles in folksonomy, Neural Netw., vol. 58, no. Supplement C, pp. 98–110, Oct.
[13] Z.-K. Zhang, T. Zhou (2010). and Y.-C. Zhang, Personalized recommendation via integrated diffusion on user–item–tag tripartite graphs, Phys. Stat. Mech. Its Appl., vol. 389, no. 4, pp. 179–186, Jan.
[14] T. Iwata, J. R. Lloyd, and Z. Ghahramani (2016). Unsupervised Many-to-Many Object Matching for Relational Data, IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 3, pp. 607–617, Mar.
[15] Kennedy, J., and Eberhart, R. C (1997). A discrete binary version of the particle swarm algorithm. In Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation., 1997 IEEE International Conference on, vol. 5, IEEE, pp. 4104–4108.
[16] Rao, R. Jaya (2016): A simple and new optimization algorithm for solving con- strained and unconstrained optimization problems. International Journal of Industrial Engineering Computations 7, 1, 19–34.
[17] Guyon, I., and Elisseeff, A (2003). An introduction to variable and feature selection. Journal of machine learning research 3, Mar, 1157–1182.
[18] Stevens, J. P (2012). Analysis of covariance. In Applied Multivariate Statistics for the Social Sciences, Fifth Edition. Routledge, pp. 299–326.
[19] Wold, S., Esbensen, K., and Geladi, P (1987). Principal component analysis. Chemometrics and intelligent laboratory systems 2, 1-3, 37–52.
[20] Schölkopf, B., Smola, A., and Müller, K.-R (1997). Kernel principal component analysis. In International Conference on Artificial Neural Networks, Springer, pp. 583–588.
[21] Zheng, W., An, L., and Xu, Z (2013). Dimensionality reduction by combining category information and latent semantic index for text categorization. JOURNAL OF INFORMATION & COMPUTATIONAL SCIENCE 10, 8, 2463–2469.
[22] Mikolov, T., Chen, K., Corrado, G. & Dean, J (2013). Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [cs].