Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 6
ISET 2024

Prediction of the popularity of artificial intelligence short videos based on MFMA

Jinbing Ha*, Yamei Gao

Faculty of Economics and Management , Nanjing University of Science and Technology , NanJing, China

*Corresponding author: 1802772568@qq.com

Abstract. In this paper, a short video popularity prediction model MFMA is designed based on Multilayer Perceptron (MLP) and artificial intelligence, and the popularity prediction in saturation is modeled as a regression problem. Through multimodal fusion, the four video attributes of visual, auditory, text, and social features are compressed as the input of the network to the greatest extent. Using nMSE and SRC as evaluation indicators, compared with other classical regression algorithms and evaluating the impact of different modal deletions on the performance of the model under the assumption of conditional independence, a large number of experiments on the real dataset of Douyin show that the MFMA model has the best prediction effect and is robust to internal noise and external uncertainty, among which the social mode has the most significant impact on the model performance.

Keywords: multimodal learning; prevalence prediction; multilayer perceptrons; deep learning.

1. Introduction

Short video popularity prediction can help service providers manage resources such as network optimization[1], provide users with satisfactory personalized recommendations[2], and bring inspiration to creators to publish high-quality content, but it also faces many challenges: (1) Heterogeneity: Short videos contain multiple elements such as images, dubbing, and title descriptions, and it is still difficult to effectively extract and integrate multimodal features. (2) Internal noise: Poor audiovisual quality of videos can be caused by few platform restrictions, lack of creator expertise, or suboptimal equipment[3]. Also, some creators use irrelevant titles and hashtags, or even fake profiles. (3) External uncertainty: the interaction and information dissemination of social networks are complex, and the platform promotion and search mechanism lead to strong randomness of popularity. These challenges make it difficult to predict the popularity of short videos.

The significance of this study is as follows: (1) The proposed MFMA model of Multi-modal feature fusion (MLP) ArtificialNeuralNetwork (ANN) can flexibly deal with internal noise and external uncertainty and make full use of the complementary information of multiple modalities. (2) The constructed short video dataset of artificial intelligence topics can be used for other related tasks. (3) Through the verification of existing feature extraction methods, the efficient extraction of a large number of video features is realized, and the similarities and differences of various features are understood.

The rest of the section is structured as follows: section II reviews the work; Section 3 elaborates on the MFMA model; Section 4 discusses the implementation of the model, including the acquisition of input features, the setting of experimental parameters, and the analysis of experimental results. Section V is concluded.

2. Related Work

This section will introduce the work in the field of multimodal learning and popularity prediction.

2.1. Multimodal learning

The research on intelligent services based on unimodal data has received extensive attention [4-5], but the information provided by unimodal data has very limited effect on the expression and dissemination of short videos, and the credibility of the prediction results is weak. Liu et al. [6]integrated the text, audio, and video features on the Vine short video platform, and established an EASTERN model for short video feature extraction and classification based on the CNN-LSTM network. Miech et al. [7] proposed a LOUPE classification model using the Youtube-8M dataset, which extracted the image and audio features of the video respectively, and analyzed the final classification through model fusion. Neural network models commonly used in general fields, such as C3D net [8] and I3D[9], are used for multimodal fusion feature extraction and analysis of video. Chen et al. [10] extracted and fused a rich set of textual, user, and time-series information to predict the popularity of social media content.

2.2. Prevalence projections

Most of the popularity predictions are for social platforms such as Twitter, Facebook, and Weibo, and there are few studies on the prediction of short videos. Zayats and Ostendorf[11] extracted text features from comments and combined user interaction graphs and long short-term memory (LSTM) networks to simulate the effects of user connection and temporal evolution on review popularity. Roy et al.[12]designed a transfer learning algorithm to analyze the impact of social trends on video popularity. Gelli et al. [13] and Trzcinski et al.[14]used support vector machine regression algorithms to predict the popularity of various images and videos posted on social platforms. Peng et al. [15]used a multi-view adaptive regression algorithm to predict the popularity of image content on social platforms.

The above-mentioned research on multimodality and prevalence prediction mainly focuses on feature engineering techniques, using only single or partial modal prediction, or using relatively simple features as social cues, and the dominant role of online content dissemination in social media is ignored. The MFMA model proposed in this study comprehensively analyzes the four modes of vision, hearing, text and society, and is robust to internal noise and external uncertainty, and can still work effectively in the face of modal loss.

3. Methodology

3.1. Problem Formulation

Given a short video dataset of with the form$\{(x_1, y_1), \dots\dots, (x_N, y_N)\}$, x_i=$\left\{x_i^{(1)}, \dots\dots, x_i^{(k)}\right\}$represents the k modal features extracted from the ith short video, the modal feature dimension of each short video is D, the features of the jth modal can be represented as $x^{(j)} \in \mathbb{R}^{\mathrm{D}\times\mathrm{N}}$, and the feature vector of the ith short video sample is $x_i = \left\{x_i^{(j)} \in \mathbb{R}^D\right\}_{\square}^{K}$, y_i is the true value of the popularity of the corresponding short video.

$$\min_{\mathrm{f}} \mathcal{L}\left(\hat{\mathrm{y}}_{\mathrm{i}} = \mathrm{f}\left(\mathrm{x}^{(\mathrm{i})}\right), \mathrm{y}_{\mathrm{i}}\right) \tag{1}$$

The goal is to learn a multi-layer perceptron model$f\left(\mathbf{x}^{(i)}\right)$, output$\hat{y}_i$, and continuously adjust the model through training, minimizing the prediction loss function $\mathcal{L}$.

3.2. Model

Compared with traditional machine learning methods, the artificial neural network based on MLP overcomes the weakness of the perceptron that cannot recognize linear and indivisible data, and can

automatically identify and select variables, avoid the influence of noise in the independent variables to the greatest extent, and improve the prediction accuracy. In this paper, a multi-layer perceptron artificial neural network model MFMA based on multi-modal feature fusion is established by using its feature extraction ability.

3.2.1. Multilayer perceptron algorithms

3.2.1.1 Basic structure

MLP generally consists of three or more layers of nonlinear activation nodes that are fully connected (an input layer, one or more intermediate hidden layers, and an output layer). In Figure 1, the input eigenvector is x_i,the hidden element of the hidden layer is h_i, o_i is the output of the output layer.

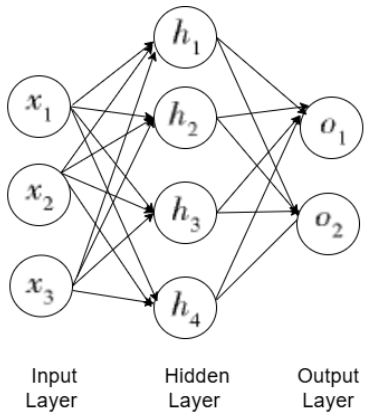

Fig. 1 Multi-layer perceptron structure

3.2.1.2 Algorithms

Forward propagation algorithm: The input data is passed to the input layer node, and the output result is obtained through hidden layer calculation and nonlinear transformation:

$$h^{(l)} = \sigma\left(W^{(l)}h^{(l-1)} + b^{(l)}\right) \tag{2}$$

$h^{(l)}$ represents the output of layer *l*, σ is the activation function, and $W^{(l)}$ and $b^{(l)}$ are the weight matrix and bias of layer *l*.

Backward propagation algorithm: The gradient of the model parameters is calculated by the loss function, and the gradient descent algorithm is used to update the parameters, so as to improve the model performance.

Specific steps:

(1) Forward propagation: The intermediate calculation results are saved during the training process for subsequent use.

(2) Calculate the loss function: Calculate the loss function according to the output result and the real label of the model.

(3) Backpropagation: The chain rule is used to calculate the gradient of the loss function for each parameter layer by layer starting from the output layer. First, calculate the error of the input layer:

$$\delta^{(L)} = \frac{\partial \mathcal{L}}{\partial z^{(L)}} \tag{3}$$

$\mathcal{L}$ is the loss function, $\delta^{(L)}$ is the derivative of the loss function with respect to the input of the output layer, and $z^{(L)}$ is the input of the output layer.

For each layer of the hidden layer $l = \mathrm{L}-1, \mathrm{L}-2, \ldots, 1$, the error of the previous layer is multiplied by the input of the current layer and the derivative of the activation function by the chain rule to obtain the error of the current layer:

$$\delta^{(\mathrm{l})} = (\mathrm{W}^{(\mathrm{l}+1)})^{\top}\delta^{(\mathrm{l}+1)} * \sigma'(\mathrm{z}^{(\mathrm{l})}) \tag{4}$$

$W^{(l+1)}$ is the weight of the layer $l+1$, $\sigma'(z^{(l)})$ is the derivative of the activation function, and $*$ represents the Hadamard product. Based on the error and the input of the current layer, the gradient of the parameters is calculated:

$$\frac{\partial \mathcal{L}}{\partial \mathrm{W}^{(\mathrm{l})}} = \delta^{(\mathrm{l})}\left(\mathrm{h}^{(\mathrm{l}-1)}\right)^{\top}, \frac{\partial \mathcal{L}}{\partial \mathrm{b}^{(\mathrm{l})}} = \delta^{(\mathrm{l})} \tag{5}$$

where $\mathrm{b}^{(\mathrm{l})}$ is the bias of layer l.

Depending on the gradient and learning rate of the parameters, the parameters of the model are updated using the gradient descent algorithm or its variants:

$$\mathrm{W}^{(\mathrm{l})} \leftarrow \mathrm{W}^{(\mathrm{l})} - \eta\frac{\partial \mathcal{L}}{\partial \mathrm{W}^{(\mathrm{l})}}, \mathrm{b}^{(\mathrm{l})} \leftarrow \mathrm{b}^{(\mathrm{l})} - \eta\frac{\partial \mathcal{L}}{\partial \mathrm{b}^{(\mathrm{l})}} \tag{6}$$

where η is the learning rate.

(4) Parameter update: According to the gradient of the parameter by the loss function, the parameters of the model are updated by using the gradient descent algorithm or its variants.

(5) Repetitive training: Repeat the previous steps until the pre-set conditions are reached (such as reaching the maximum number of iterations, loss function convergence, etc.).

The MLP model established in this paper consists of three layers: one input layer, three hidden layers, and one output layer. The activation function uses the Relu function, and the weights and biases of the model are calculated by the backward propagation algorithm.

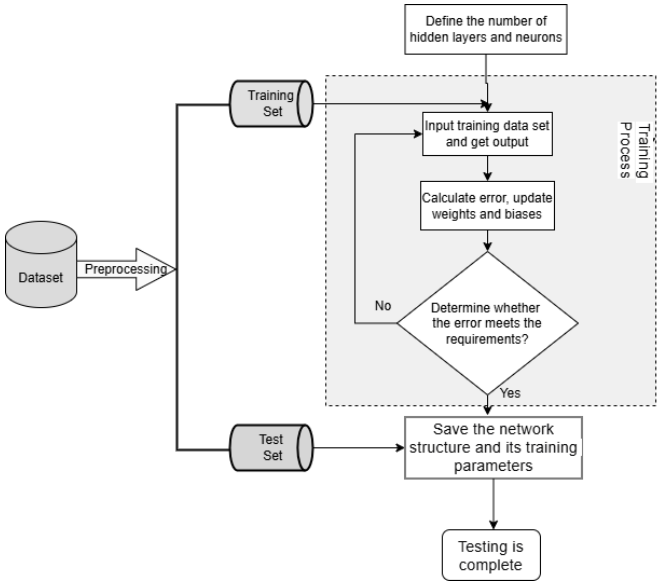

Fig.2 Flowchart of model building

3.2.2 Feature extraction

This section describes the feature selection and extraction methods for the four modes of short videos.

3.2.2.1 Visual features

The fixed time interval method is used to select the keyframes, and then the features extracted from the keyframes are fused through the average pooling operation.

(1) Global features: The color space is divided into 64 types, and a 64-dimensional color probability histogram feature vector is obtained for each frame of the image, representing the basic visual features and describing the color style and theme of the short video [16].

(2) Object features: Using the AlexNet network trained on the Caffe platform [17], the keyframe image of the short video sample is input, and the 1000-dimensional vector (the class score of the 1000 types of objects predefined in the Image Net dataset) before the final softmax classification layer is selected as the target type feature of the short video.

(3) Sentiment features: DeepSentiBank, a deep CNN visual sentiment analysis model trained on the SentiBank [18] dataset, selected 2089 different ANPs to describe the sentiment features in the model.

3.2.2.2 Auditory characteristics

The Mel Frequency Cepstral Coefficient (MFCC) [19] and six basic sound frequency features (energy entropy, signal energy, zero-crossing rate, spectral roll point, spectral center, and spectral flux) were extracted from the audio channels of short video samples using Python's libsora audio feature extraction library.

3.2.2.3 Text features

In this paper, the semantic features used for content description and the emotional features used for sentiment description are extracted.

(1) Semantic features: Based on the text feature extraction tool MiniRBT-H256 [20], Chinese pre-training is carried out by using knowledge distillation technology and whole word masking technology, and finally a 256-dimensional vector description of text semantic features is obtained.

(2) Affective characteristics: Berger[21] has shown the influence of user emotion on behavior. Using the text sentiment analysis tool in SnowNLP, a natural language processing toolkit, the text sentiment information is described as a 1-dimensional vector, and the value from 0 to 1 represents the probability of positive emotion.

3.2.2.4 Social characteristics

In addition to video content, the popularity of short videos is also affected by the creator's personal information. "Influencer" or platform-certified "professional" accounts with a strong fan base tend to post videos with higher exposure and are more likely to gain higher popularity. In this article, five indicators of the creator's number of fans, the number of followers, the total number of likes, the total number of works, and whether the creator is certified on the Douyin platform represent the social characteristics.

3.2.2.5 Summary of feature models

Through early fusion, the four characteristics are fused to obtain a comprehensive feature set for short video popularity prediction:

Table 1. Multimodal characteristics of short videos

modality	name of the feature	dimension
visual	Color histogram	64
	Target characteristics	1000
	Emotional traits	2089
audio	Mel frequency cepstrum coefficient and Six sound frequency characteristics	19
text	Textual semantic features	256
	Sentiment characteristics of the text	1
social	up_follower	1
	up_following	1
	total_favor	1
	aweme_cnt	1
	Platform Certification	1

The feature extraction process is done offline and does not increase the computational complexity of subsequent models. In addition, some of the eigenvectors have a large range of values, so all eigenvectors are normalized based on the L2 norm.

3.2.3 Short video popularity prediction model MFMA

Figure 3 shows the overall model framework for predicting the popularity of artificial intelligence short videos based on the MFMA model.

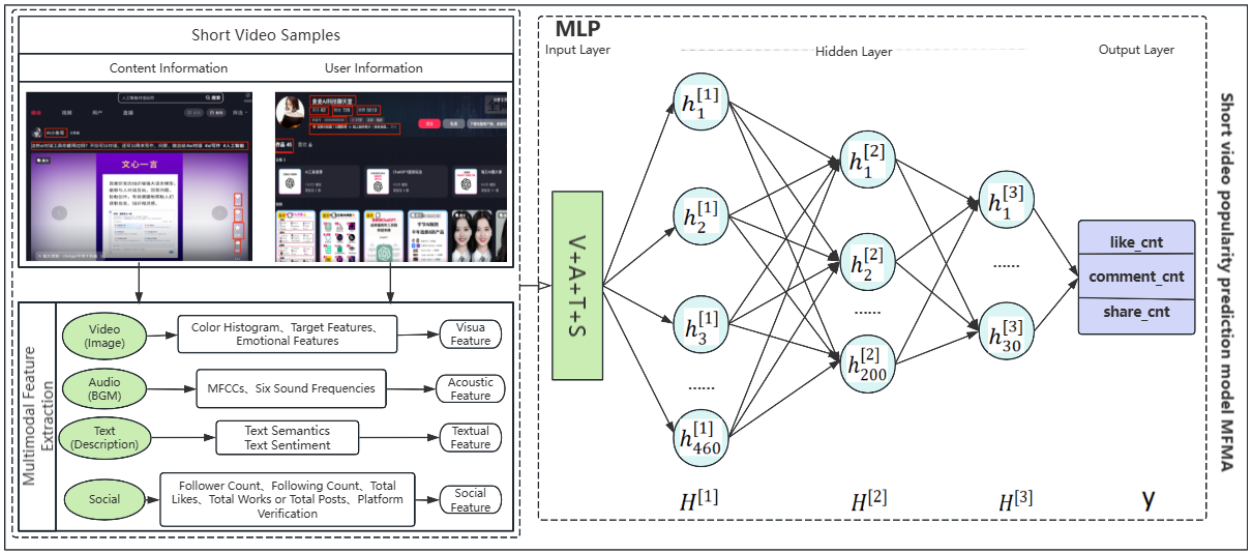

Fig .3 Overall model framework

The CPU configuration of the computer used for model building and training is Intel(R) Core(TM) i7-6500U CPU @ 2.50GHz 2.60 GHz, 16 GB memory, the model is implemented by Keras[22]and sklearn[23], and the training is accelerated by Tensorflow GPU, the programming language is Python 3.9, and the operating system is Windows 10.

4. Experiments

4.1. Datasets

The data comes from the Douyin platform. First, use the downloader to crawl the target video information, and then use the Python crawler to call the Douyin API to obtain the video author information according to the author_id corresponding to the video_id. The final target dataset is shown in Figure 4:

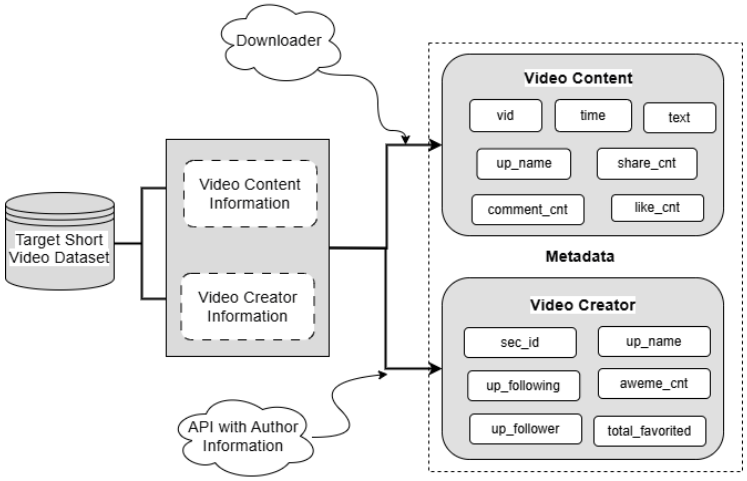

Fig. 4 Flowchart of dataset collection

Table 2 describes the attributes of the dataset:

Table 2. Attributes of the short video dataset

Attribute Name	Description	Attribute Name	Description
Video content information		Creator information	
vid	Video id	sec_id	Author id
up_name	Nickname of the author	up_name	Nickname of the author
text	Video title, description	aweme_cnt	The total number of works by the author
time	The length of the video	up_following	Number of followings
like_cnt	Likes	up_follower	Number of followers
comment_cnt	Number of comments	total_favor	The total number of likes the author received
share_cnt	Number of forwards		

The Douyin platform limits the number of searches, so synonym substitution is used to formulate short video keywords related to artificial intelligence (artificial intelligence, machine learning, reinforcement learning, expert systems, etc.) [24]. The dataset contains 9,832 short video samples from September 20, 2016 to March 17, 2024.

To define the popularity of short videos:

$$y_i = \log_{10}[\text{like\_cnt}_i, \text{comment\_cnt}_i, \text{share\_cnt}_i] \quad (7)$$

where y_i is the true value of the popularity of the ith video,$like\_cnt_i, comment\_cnt_i, share\_cnt_i$ which is the number of likes, comments, and retweets of the ith video after it was published for at least half a month[25].

In order to ensure the consistency of the scale and scope of the data, and improve the training effect and generalization ability of the model, the Z-score standardization method was selected to process the modal feature data and popularity score label data of the input layer:

$$x_{normalized} = \frac{x-\mu}{\delta} \tag{8}$$

x is the original eigenvalue, μ is the eigenmean, and δ is the eigenstandard.

4.2. Experimental Setup

20 rounds of random experiments were carried out on the dataset, 80% of the samples were used as the training set, 20% were used as the test set, and the average results of the test were finally obtained. The training set is sent to the built network, and the 0.1 training set is divided into verification sets during the training process. The optimal model was selected through multiple experiments and parameter tuning: there were 3 hidden layers, the number of neurons in each layer was 460, 200, and 30, respectively, and the output layer was only 1 neuron, and Relu was used as the activation function. In the training process, Adam dynamically adjusts the learning rate to accelerate the convergence speed of the model, and uses the early stopping method to terminate the training process if necessary. The Stochastic Gradientdescent (SGD) algorithm was used to optimize, and the learning rate was set to 0.01 according to the experience.

4.3. Prevalence Predictions

4.3.1. Evaluation indicators

The agreement between the predicted and true values is measured by the normalized mean squared error (nMSE)[26-27]:

$$nMSE = \frac{1}{N\sigma^2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{9}$$

N is the number of short video samples, y_i and $\hat{y}_i$ are the true and predicted values of short video popularity, and σ is the standard deviation of the true value of short video popularity. The lower the nMSE value, the better the model performance.

The Spearman's Rank Correlation (SRC) [28]is used as a supplementary measure to measure the nonlinear relationship between the two variables:

$$SRC = 1 - \frac{6\sum d_i^2}{N(N^2-1)} \tag{10}$$

d_i^2 represents the D-value of the predicted popularity and true popularity series of the ith short video, indicating the similarity between the predicted and true popularity sequences[29]. The higher the SRC value, the better the model performance.

4.3.2. Results and Analysis

(1) Convergence analysis

An experiment was randomly selected to observe the trend of nMSE with the number of iterations:

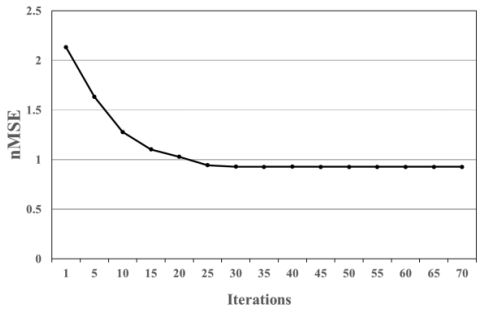

Fig.5 Convergence curve

As can be seen from Figure 5, nMSE gradually decreases with the increase of iterations and tends to converge to about 25 iterations, which verifies the feasibility of the algorithm.

(2) the lack of robustness in different modalities

Under the assumption of conditional independence, experiments are carried out on several trained models with different data segmentation in the face of new samples with fewer modalities, and the results are shown in Table 2:

Table 2. Performance of MFM3 in modal loss

Modal combinations	nMSE	SRC
T+V+A+S	*0.947*	*0.774*
T+V+A	0.962(+0.015)	0.690 (-0.084)
T+V+S	0.952(+0.05)	0.736(-0.038)
T+A+S	0.953(+0.06)	0.735(-0.039)
V+A+S	0.949(+0.02)	0.770(-0.004)

Overall, removing any one modality results in reduced performance due to the different modalities having different levels of expression. However, the impact is negligible, as the intrinsic correlation of the different modalities may complement each other. Table 2 also shows that different modalities have different effects on model prediction, and the absence of social modality has the most significant impact on performance, because the number of followers and the number of authors' followers in social modality reflect the creator's influence and credibility on the Douyin platform, which directly or indirectly affects the popularity of videos. In addition, the elimination of visual and auditory modalities has a similar impact on model performance; The absence of text modality has the least impact on model performance, because many short video samples lack text descriptions or are caused by weak correlation between short videos and text.

Sorted in descending order of modality importance to the predictive model: social > visual> audio > text. The model has the best performance when the four modalities are combined, which indicates that the MFMA model is effective in using the complementary information of different modal features.

(3) Comparison between the MFMA model and the benchmark method
In order to further verify the effectiveness of the model, the MFMA model is compared with the benchmark method:

General multiple linear regression(MLR): Linear equations are used to capture the dependencies between two or more independent and dependent variables for modeling, which is an extension of classical linear regression[30].

Ridge regression: The applicability to pathological data is better than the least squares method[31], and the Bayesian Ridge regression algorithm is used for ridge regression in this paper.

Support vector machine regression (SVR): In this study, the Gaussian radial basis function kernel (RBF kernel function) in SVR was used as a regression method with a maximum margin criterion [32].

Lasso regression: The principle is to reduce the dimensionality by changing the weight of certain features to zero, which is a type of regularization of regression algorithms [33].

Table3. Performance comparison of the MFM4 model with the benchmark method

method	nMSE	SRC
MLR	1.132	0.370
Bayesian Ridge returns	0.982	0.736
SVR-RBF	0.979	0.772
Lasso returns	0.975	0.801
MFMA	*0.893*	*0.831*

The results show that the multiple regression model may not be able to match the nonlinear relationship between the popularity and characteristics of short videos based on the linear relationship, so the prediction performance is relatively poor. Although the radial basis function kernel support vector machine regression can deal with the nonlinear problem, the selection of the kernel function and the adjustment of parameters are relatively complex, which may lead to the poor performance of the model. Although Lasso regression can be used for feature selection and prevent overfitting, it has certain limitations, and it is difficult to fully capture the complex relationship between the popularity of short videos and multimodal features.

The MFMA model can fully learn multimodal features and use the nonlinear fitting ability of MLP to capture the complex relationship between short video popularity and modal features, so the prediction performance of the model is higher than that of other classical machine learning regression algorithms.

5. Summary

The short video prediction model MFMA constructed in this study uses the complementary information of multiple modalities to analyze the short video picture, audio, text and creator's social characteristics to predict the popularity of short videos, and evaluate the dissemination ability of artificial intelligence short videos on the Douyin platform. Achieve efficient extraction of a large number of video features. A large number of experiments have proved the effectiveness of the MFMA model. In future work, the influence of social features on information diffusion can be further explored, which may further improve the interpretability and performance of predictions.

References

[1] M. Zeng et al., “Temporal-spatial mobile application usage understanding and popularity prediction for edge caching,” IEEE Wireless Commun.,vol. 25, no. 3, pp. 36–42, Jun. 2018.

[2] A. Bielski and T. Trzcinski, “Understanding multimodal popularity pre-diction of social media videos with self-attention,” IEEE Access,vol.6,pp. 74277–74287, 2018, doi: 10.1109/ACCESS.2018.2884831.

[3] S.Wang et al.,“Video affective content analysis by explor-ing domain knowledge,”IEEE Trans. Affective Comput.,doi:10.1109/TAFFC.2019.2912377.

[4] Weifan Xie, Yan Guo, Guang sheng Kuang, et al. Prediction method of Twitter topic popularity based on evolutionary model[J].Computer Applications,2022,42(11):3364.

[5] Ding J, Li Y, Li Y, et al. Click versus share: A feature-driven study of micro-video popularity and virality in social media[A]//Proceedings of the 2018 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics[C], 2018:198-206.

[6] MiechA, LaptevI, SivicJ. Learn ablepooling with context gating for video classification [J]. arXiv preprint arXiv: 1706.06905,2017.

[7] TranD, BourdevL, FergusR, et al. Learning spatio temporal features with 3d convolutional networks[C]//Proceedings of the IEEE international conference on computer vision. 2015: 4489-4497.

[8] CeronA, CuriniL, IacusSM, et al. Every tweetcounts? How sentiment analysis of social media can improve our knowledge of citizens’ political preferences with anapplication to Italy and France[J]. New media&society, 2014,16(2):340-358.

[9] GoesPB, LinM, Au YeungC. “Popularityeffect”inuser-generated content: Evidence from online product reviews[J]. Information Systems Research, 2014, 25(2): 222-238.

[10] RoyS D, MeiT, ZengW,et al. Towards cross-domainl earning for social video popularity prediction[J]. IEEE Transactions on multimedia, 2013,15(6): 1255-1267.

[11] V.Zayats and M.Ostendorf,“Conversation modeling on using a graph-structured,”Trans.Assoc. Comput. Linguistics, vol.6, pp.121–132, 2018.

[12] PengH, LiK, LiB, et al. Predicting image memor ability by multi-view adaptive regression[C]//Proceedings of the23rd AC Minternational conference on Multimedia. 2015:1147-1150.

[13] GelliF, UricchioT, BertiniM, et al. Image popularity predictionin social media using sentiment and context features[C]//Proceedings of the 23rd ACM international conference on Multimedia.2015: 907-910.

[14] TrzcinskiT, RokitaP. Predicting popularity of online videos using support vector regression[J]. IEEE Transactions on Multimedia, 2017, 19(11): 2561-2570.

[15] PengH, LiK, LiB, et al. Predicting image memor ability by multi-view adaptive regression[C]//Proceedings of the 23rd ACM international conference on Multimedia. 2015: 1147-1150.

[16] Hendrycks D, Gimpel K. Gaussian error linear units (gelus)[J]. arXiv preprint arXiv:1606.08415, 2016.

[17] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems, 2012, 25.

[18] Borth D, Chen T, Ji R, et al. Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content[A]//Proceedings of the 21st ACM international conference on MultimediarC].2013:459-460.

[19] Borth D, Chen T, Ji R, et al. Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content[A]//Proceedings of the 21st ACM international conference on MultimediarC].2013:459-460.

[20] Jiao X, Yin Y, Shang L, et al. Tinybert: Distilling bert for natural language understanding[J]. arXiv preprint arXiv:1909.10351, 2019.

[21] Berger J. Arousal increases social transmission of information[J]. Psychological Science, 2011, 22(7): 891-893.

[22] Galea A, Capelo L. Applied Deep Learning with Python: UseScikit-learn, TensorFlow, and Keras to Create IntelligentSystems and Machine Learning Solutions. Packt Publishing,2018.

[23] Hackeling G. Mastering Machine Learning with Scikit-learn.2nd ed. Packt Publishing, 2017.

[24] Liu. Research on short video prevalence prediction model of cardiovascular disease topic based on deep learning[D]. Huazhong University of Science and Technology, 2021. DOI:10.27157/d.cnki.ghzku. 2021.005713.

[25] Roy S D, Mei T, Zeng W, et al. Towards cross-domain learning for social video popularity prediction[J]. IEEE Transactions on multimedia, 2013, 15(6): 1255-1267.

[26] J.Chen et al., “Micro tells macro: Predicting the popularity ofmicro-videos via a transductive model,” in Proc. ACM Int. Conf. Multimedia, 2016,pp. 898–907.

[27] P. Jing et al., “Low-rank multi-view embedding learning for micro-video popularity prediction,” IEEE Trans. Knowl. Data Eng., vol. 30, no. 8,pp. 1519–1532, Aug. 2018.

[28] A. Khosla, A. DasSarma, and R. Hamid,“Whatmakes an image popular?,” in Proc. Int. Conf. World Wide Web, 2014, pp. 867–876.

[29] ZHANG Zhuoran. Short video popularity prediction and value evaluation based on multimodality[D]. Beijing University of Posts and Telecommunications, 2023. DOI:10.26969/d.cnki.gbydu. 2023. 000082.

[30] LIN Bin. Multiple Linear Regression Analysis and Its Applications[J]. China Science and Technology Information, 2010, 9(02): 2011.

[31] YANG Nan. The Unique Role of Ridge Regression Analysis in Solving Multicollinearity Problems[J]. Statistics & Decision-Making, 2004 (3): 14-15.

[32] Smola A J, Schölkopf B. A tutorial on support vector regression[J]. Statistics and computing,2004, 14(3): 199-222.

[33] Kukreja S L, Löfberg J, Brenner M J. A least absolute shrinkage and selection operator (LASSO) for nonlinear system identification[J]. IFAC proceedings volumes, 2006, 39(1): 814-819.