doi:10.3772/j.issn.1006-6748.2023.01.011

Prediction of film ratings based on domain adaptive transfer learning[①]

SHU Zhan(舒　展), DUAN Yong[②]
(School of Information Science and Engineering, Shenyang University of Technology, Shenyang 110870, P. R. China)

Abstract

This paper examines the prediction of film ratings. Firstly, in the data feature engineering, feature construction is performed based on the original features of the film dataset. Secondly, the clustering algorithm is utilized to remove singular film samples, and feature selections are carried out. When solving the problem that film samples of the target domain are unlabelled, it is impossible to train a model and address the inconsistency in the feature dimension for film samples from the source domain. Therefore, the domain adaptive transfer learning model combined with dimensionality reduction algorithms is adopted in this paper. At the same time, in order to reduce the prediction error of models, the stacking ensemble learning model for regression is also used. Finally, through comparative experiments, the effectiveness of the proposed method is verified, which proves to be better predicting film ratings in the target domain.

Key words: prediction of film rating, domain adaptive, transfer component analysis (TCA), correlation alignment (CORAL), stacking

0 Introduction

With the development of machine learning, more and more algorithms have been widely applied in daily life, including but not limited to speech recognition, language translation, target detection and weather prediction. As film data have the features of being in volume, various types, and coming from wide varieties of sources, how to use these data to help producers and investors with inspirations in their future production processes, and to even promote the development of film industry is an urgent problem[1].

However, the training dataset and testing dataset used in practical application scenarios mostly follow different marginal distributions, and the traditional machine learning model requires a large number of data samples with complete sample labels to participate in the training process. However, this dataset cannot be completely obtained for many reasons, which leads to the difficulty in training. Transfer learning refers to a learning process that applies the knowledge learnt in the old domain to a new domain by using the similarity between data, tasks or models[2-4]. Domain adaptation, a sub-research field in transfer learning, is defined as a scene that the source domain and the target domain share the same features and categories but have a different feature distribution. Target domain data are calibrated by using the source domain data with rich annotation information[5]. In order to solve the problem where the public dataset of pedestrian attribute identification is not easy to be directly applied to real life, Cheng et al.[6] proposed a multi-source and multi-label pedestrian attribute recognition method based on domain adaptive approaches, which can effectively solve the problem of feature heterogeneity in multiple datasets, and the average prediction accuracy is also improved. Noh et al.[7] proposed a domain adaptation method based on cycle-consistent generative adversarial networks(CycleGAN) to solve the problem of performance degradation of finger vein recognition model trained in one dataset to predict additional datasets, which can effectively improve the prediction accuracy on data from the target domain.

Taking Douban film dataset as the source domain and Internet movie database (IMDB) as the target domain, this paper studies how to use domain adaptive algorithm to solve the unsupervised transfer learning and to effectively predict the film ratings in the target domain. In the feature engineering, based on original features from the dataset, new features, such as levels of actors and directors, and composite type influence index are constructed. For reducing the influence of singular film samples in model training, they are re-

[①] Supported by the Scientific Research Foundation of Liaoning Provincial Department of Education (No. LJKZ0139).

[②] To whom correspondence should be addressed. E-mail: duanyong0607@163.com.

Received on Dec. 8, 2021

moved through the clustering algorithm. In order to find the best prediction features from many candidates, the film features are also selected. Because feature dimensions of film samples in the source domain and the target domain are inconsistent, this paper applies the dimensionality reduction algorithm to align the feature dimensions of data from source domain and target domain, and uses stacking ensemble learning model as a regression transfer component analysis (TCA) and correlation alignment (CORAL) domain adaptive transfer learning model in the study of film rating predictions in the target domain.

1 Feature engineering for film data

The source domain dataset used in this paper is the film dataset provided by Douban, which is a Chinese film review website, and the target domain dataset is provided by IMDB, which is an American film review website. These datasets contain director names, actor names and other features of each film. At present, researchers take these datasets as the object to carry out studies on machine learning or data analysis in film industry.

1.1 Features of composite type influence index for constructing film samples

The selected original features of the dataset are box office, budget, rating, the number of user reviews, the number of film review experts, the number of voters and the number of Facebook fans so that these features are used to retrieve a weighted sum. The average value of the final sum is then used as the influence index of the film type, as shown in Eq. (1).

$$\lambda_c = \frac{\sum_{i=1}^{n_c} \sum_{b=1}^{7} \gamma_{ib} a_{ib}}{n_c} \tag{1}$$

where, λ_c is the influence index of film type c, n_c is the number of samples containing film type c in the film attribute, a_{ib} is the selected feature of the film sample i, and γ_{ib} is the weight value corresponding to each feature.

Usually, a film often has a variety of film type features, which can be summed according to the corresponding influence index with respect to the film subject type processed by the one-hot coding above and to construct the composite type influence index of each film sample $genres\_influ_n$, as shown in Eq. (2).

$$genres\_influ_n = \sum_{c=1}^{m} \lambda_c \tag{2}$$

where, $genres\_influ_n$ is the composite type influence index of the nth film sample, m is the number of film types contained in each film, and λ_c is the influence index of film type c.

1.2 Features of directors and actors level for constructing film samples

To construct the level features of directors and actors, the K-Means ++ algorithm is used in this paper. Compared with the traditional K-Means clustering algorithm, K-Means ++ improves the selection of initial points, so that the distance between the initial clustering centers can be as large as possible.

For the determination of K value, this paper uses grid-search method combined with contour coefficients and sum of squares for error (SSE) coefficients. The K value in the process is four for directors and three for actors. The directors and actors will have a new feature K-Means _ label to mark the cluster corresponding after clustering. It can be referred to the average value of each feature contained in every cluster sample to find out which cluster contains more features of directors or actors, and then re-integrates the K-Means _ label features. And, the larger the eigenvalue is, the better each feature of the corresponding sample in the cluster is. After these steps, each director and actor can get a level feature and attach to original film dataset for constructing the director _ label, actor _ 1 _ label, actor _ 2 _ label and actor _ 3 _ label features for each film sample.

1.3 Processing of singular film samples

Based on the existing sample data, in order to improve the accuracy of the model prediction, some singular film samples can be removed. The preferred method for removing singular samples is to cluster the remaining film samples, judging from the Euclidean distance between each film sample and the center of its cluster. Decisions for removing them can be made according to a set threshold.

The clustering algorithm used in this step is hierarchical clustering with four being the number of clusters. Since it is impossible to identify the center of each cluster through hierarchical clustering, it is necessary to calculate the clustering again. The mean values of every feature of the film samples contained in each cluster are combined as the center of each cluster. For each cluster, the Euclidean distance between the sample contained in the cluster and the corresponding cluster center should be obtained. The threshold for removing a singular film sample is set to the average Euclidean distance between the sample contained in each cluster and its cluster center.

1.4 Selecting film sample features

Since not all of film sample features are helpful to the film ratings prediction, most of feature selection methods rely on the tree model. As in the process of constructing a single decision tree, the splitting of sub-nodes is based on the information entropy, the conditional entropy and the information gain[8]. Information entropy is an indicator used to measure information uncertainty, as shown in Eq. (3).

$$H(X) = -\sum_{i=1}^{n} P(X = i) \log_2 P(X = i) \qquad (3)$$

where, $H(X)$ is the information entropy, $P(X = i)$ is the probability of a random variable X when it is equal to i.

Conditional entropy represents the uncertainty of random variable X under the condition of given random variable Y, as shown in Eq. (4).

$$H(X \mid Y) = -\sum_{i=1}^{n} P(X = i \mid Y) \log_2 P(X = i \mid Y) \qquad (4)$$

Information gain represents the degree of information uncertainty reduction under certain conditions, and it can be calculated as

$$I(X,Y) = H(X) - H(X \mid Y) \qquad (5)$$

where, $I(X,Y)$ denotes the information gain, $H(X)$ is the information entropy, and $H(X \mid Y)$ is the conditional entropy.

When constructing a decision tree, features with large information gain caused by feature splitting are selected as the segmentation features. Therefore, important features are more likely to appear near the root node, while unimportant features usually appear near the leaf nodes or do not appear at all[9]. In this paper, the extreme gradient boosting (XGBoost) model based on multiple decision trees is utilized to filter features according to the average information gain brought by feature splitting. After removing the features having lower importance ranking, 38 remaining features can be kept for the subsequent model training and the film rating prediction processes.

2 Film rating prediction model based on domain adaptive transfer learning

After the feature engineering is completed, as the film datasets in the source domain and the target domain come from different websites, the data samples obey different marginal distributions, which leads to a situation that the traditional machine learning model cannot be used for training and predicting. In order to address the problem, this paper instead uses the domain adaptive transfer learning model.

Currently, domain adaptive transfer learning algorithms can be divided into three categories. From the perspective of data distribution, if the probability distributions of data samples in the source domain and the target domain are similar, it can be achieved by minimizing the probability distribution distance. From the perspective of feature selections, these features can be selected if the source domain and the target domain data samples share some common features. From the perspective of feature transformation, if the data samples of the source domain and the target domain share some subspaces, the two domains can then be transformed into the same subspace.

Based on these, this paper studies film rating prediction in the target domain under data distribution and feature transformation. The TCA algorithm based on data distribution and the CORAL algorithm based on feature transformation are used for modelling.

2.1 The TCA film rating prediction model with principal components analysis (PCA) dimension reduction

The TCA is a marginal distribution adaptive method based on maximum mean difference (MMD) distance to identify the difference of film sample data distribution in the source domain and target domain, and its target solution can be defined as Eq. (6)[10].

$$(\boldsymbol{XMX}^{\mathrm{T}} + \lambda \boldsymbol{I})\boldsymbol{A} = \boldsymbol{XHX}^{\mathrm{T}}\boldsymbol{A\Phi} \qquad (6)$$

where, $\boldsymbol{X}$ is the source domain and the target domain film sample feature matrix, $\boldsymbol{M}$ is the MMD distance matrix, $\boldsymbol{I}$ is the unit matrix, λ is the equilibrium parameter, $\boldsymbol{H}$ is the center matrix, $\boldsymbol{\Phi}$ is the Lagrange parameter, $\boldsymbol{A}$ is the transformation matrix to be solved.

The MMD distance can be defined as Eq. (7).

$$MMD(P_s(x), P_t(x)) = \left\| \frac{1}{N_s}\sum_{i=1}^{N_s} \boldsymbol{A}^{\mathrm{T}} x_i - \frac{1}{N_t}\sum_{j=1}^{N_t} \boldsymbol{A}^{\mathrm{T}} x_j \right\|_H^2 \qquad (7)$$

where, $P_s(x)$ and $P_t(x)$ represent the marginal distribution of the source domain and target domain film samples, N_s and N_t represent the number of film samples in the source domain and target domain, and x_i or x_j represent individual film samples in the source domain or target domain[11].

Firstly, when solving the feature transformation matrix $\boldsymbol{A}$, the MMD distance matrix $\boldsymbol{M}$ and the center matrix $\boldsymbol{H}$ can be obtained with respect to the source domain and target domain film sample feature matrix $\boldsymbol{X}$. Then the solution can be made through selecting the linear kernel function or Gaussian kernel function mapping. For feature transformation matrix $\boldsymbol{A}$, the first d features are generally taken to obtain the data after dimension reduction in the source domain and target domain. Finally, the stacking ensemble learning model is used to bring the film data of the current source domain into the model for training, and then the film sample ratings in the current target domain can be predicted.

For the input feature matrix $\boldsymbol{X}$ of film samples in the source domain and target domain, feature dimensions of film samples from both domains need to be consistent, otherwise the subsequent matrix operations cannot be performed. Since the Douban film dataset used in this paper is the source domain and the IMDB film dataset is the target domain, their feature dimensions are inconsistent, so it is necessary to apply the dimensionality reduction algorithm to align their feature dimensions. Feature dimensions of samples from the Douban film dataset is relatively large, and it has to be reduced to 38 to achieve consistency with IMDB. Since the adaptive transfer learning model in the TCA field is a transfer learning model from the perspective of data distributions, and the PCA dimension reduction algorithm[12] can also be used to extract main feature components in the source domain film samples from the perspective of data distributions and to retain the main feature information in the original data, so that the attribute features after dimension reduction are independent of each other. The combination can effectively solve the problem of feature dimension inconsistency and film rating prediction in the target domain.

2.2 The CORAL film rating prediction model with locally linear embedding (LLE) dimension reduction

The CORAL is a subspace transformation method for second-order feature alignment of the source domain and target domain. Let $\boldsymbol{C}_s$ and $\boldsymbol{C}_t$ be the covariance matrices of film samples in the source domain and target domain, matrix $\boldsymbol{A}$ is a second-order feature transformation matrix, which minimizes the feature distance between source domain and target domain, as shown in Eq. (8)[13].

$$\min_{\boldsymbol{A}} \| \boldsymbol{A}^{\mathrm{T}}\boldsymbol{C}_s\boldsymbol{A} - \boldsymbol{C}_t \|_F^2 \qquad (8)$$

The domain adaptive transfer learning model in the CORAL domain also involves the alignment of feature dimensions of the input film samples in the source domain and target domain. Since the CORAL model mines spatial geometric features of film samples in the source domain and target domain from the perspective of feature transformations, and the LLE is also a nonlinear dimensionality reduction algorithm that maintains the original spatial manifold structure after the dimensionality reduction of the film data from source domain[14]. Therefore, the CORAL domain adaptive transfer learning model together with the LLE dimensionality reduction algorithm can achieve better feature dimension alignment and film rating prediction.

3 The domain adaptive film rating prediction model based on stacking

The domain adaptive transfer learning model combined with specific dimension reduction algorithms is used to process the features of the film data samples from source domain and target domain, and then the stacking ensemble learning model is used for regression. The ensemble learning is to complete the learning task by constructing and combining multiple learners, and the predicted results are often better than an individual learner. As shown in Fig. 1, the general structure of ensemble learning is to generate a group of individual learners, and then combine them with a certain combination strategy. These individual learners are usually generated from the training dataset through existing learning algorithms[15].

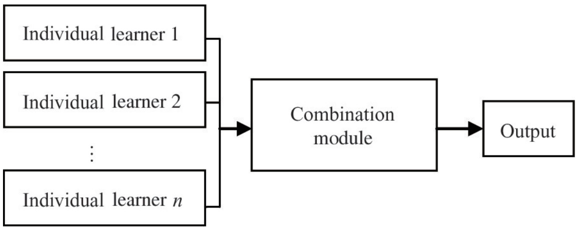


Fig. 1 Schematic diagram of the ensemble learning

At present, there are three main ensemble learning strategies: the average method, the voting method, and the learning method, and the learning method is more representative of the stacking ensemble learning method. The stacking ensemble learning model constructed in this paper has a two-layer structure. The first layer combines K-nearest neighbor (KNN), decision tree, support vector machine (SVM), random forest, XGBoost, ridge regression, extra trees, AdaBoost and gradient boosting decision tree (GBDT) models as individual learners. Each learner performs a 5-fold cross validation, and a new training dataset can be generated. In order to alleviate the over-fitting phenomenon, the second layer learner chooses linear regression model, which will be trained according to the new training dataset. In prediction, the linear regression model uses the average prediction results from each individual learner to predict film ratings in the target domain[16].

4 Experimental results and analyses

Since predicting film ratings is a typical regression problem in machine learning, some evaluation metrics for regression algorithms can be used in model evaluation, such as mean absolute error (MAE) and mean squared error (MSE). They are shown in Eq. (9) and Eq. (10).

$$MAE = \frac{1}{m}\sum_{i=1}^{m}(| y_i - f(x_i) |) \qquad (9)$$

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - f(x_i))^2 \qquad (10)$$

where, m is the number of film samples, y_i is the real rating of the ith film sample, and $f(x_i)$ is the rating of the ith film sample predicted by the model.

In this paper, Douban film dataset is selected as the source domain data, while IMDB film dataset is selected as the target domain data. Firstly, the data preprocessing and the feature engineering of these film datasets are carried out. Then the domain adaptive algorithm is studied under the premise of the target domain data without sample labels. Due to the large attribute feature dimension of Douban film data, the feature dimensions of film samples from source domain and target domain are inconsistent. Therefore, dimension reduction is used to reduce the number of dimensions to 38. When studying the domain adaptive TCA model, the PCA dimension reduction algorithm is also used to reduce the dimension of features for the film samples from source domain, and then the linear kernel function is used for mapping. Finally, in order to improve the prediction accuracy of the model and reduce the prediction error, the stacking ensemble learning model is used for regression in this paper. The MAE and MSE results from the TCA model prediction are shown in Fig. 2 and Fig. 3.

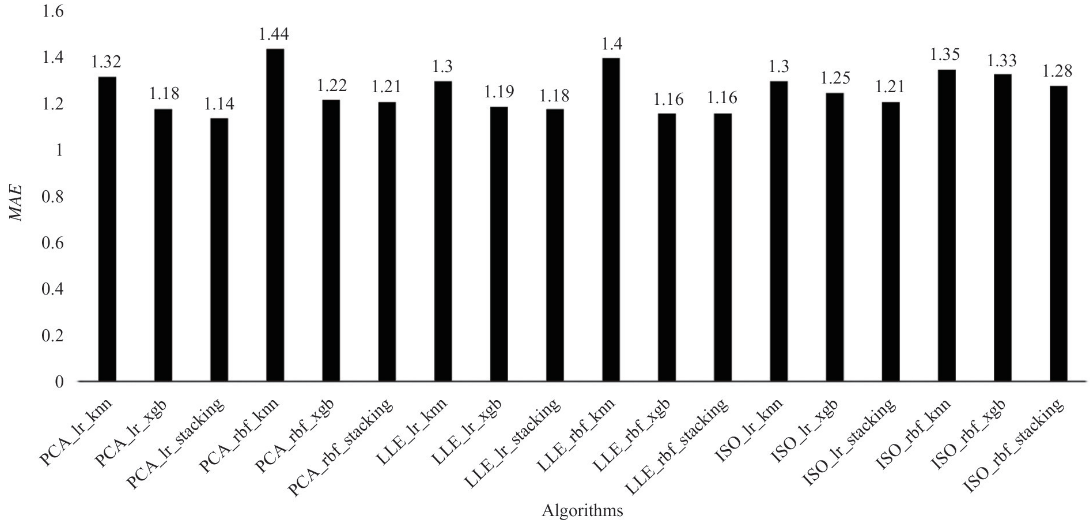


Fig. 2 The MAE of TCA model prediction results

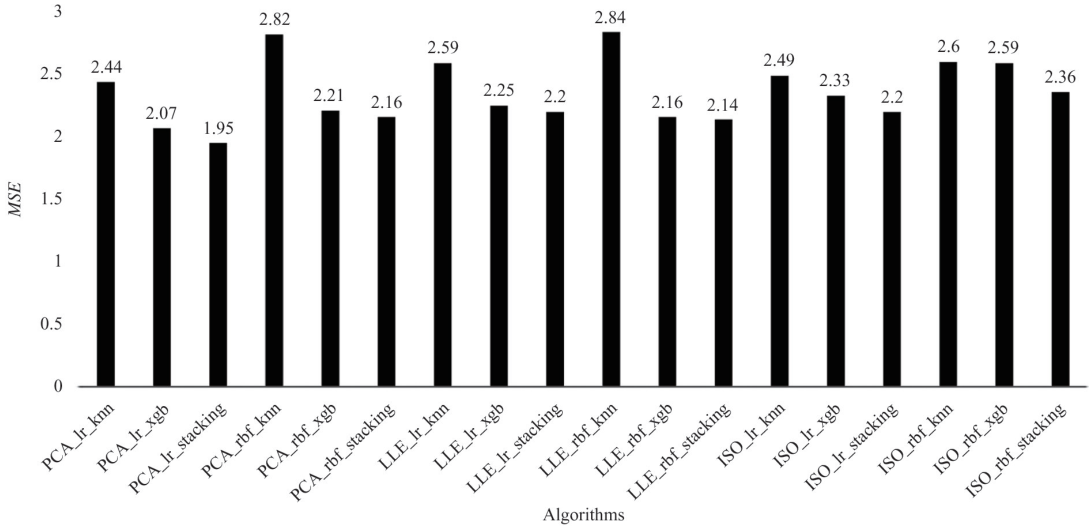


Fig. 3 The MSE of TCA model prediction results

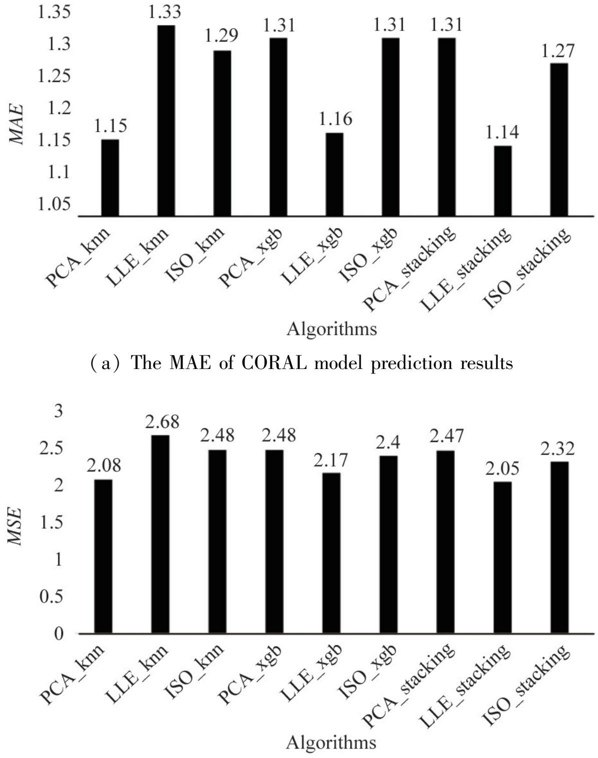
In the dimensionality reduction algorithm of the TCA model, LLE and isometric mapping (ISOMAP) are selected as the comparison objectives of PCA. The Gaussian kernel function is used in the comparison of linear kernel function mapping. In terms of regression selection, KNN and XGBoost are selected as the comparison objects of the stacking ensemble learning model. From Fig. 2 and Fig. 3, the TCA model with PCA, linear kernel function and stacking ensemble learning model for regression has the best prediction outcome, with an MAE of 1.14 and an MSE of 1.95.

Compared with the other dimensionality reduction algorithms, MAE and MSE of the model using PCA to reduce the dimensionality of the source domain film data are better, from the perspective of the mapping effect of the kernel function, the linear kernel function mapping is generally better than that of the Gaussian kernel function mapping. This is due to the fact that the feature matrix composed of the source domain film data after dimensionality reduction and the target domain film data tends to be linearly separable. From the perspective of regressors, the TCA model based on stacking ensemble learning predicts film ratings better than TCA models with KNN and XGBoost. The TCA model studied in this paper has an average relative decrease of MAE by about 9.4% and MSE by about 17.3% compared with the comparative objects.

In summary, the combination of PCA to reduce the dimension of source domain data can effectively solve the problem of inconsistent feature dimensions between the source domain and target domain. The TCA model using linear kernel function mapping and the stacking ensemble learning model for regression can effectively solve the problem of film rating prediction in the target domain, and it can achieve better prediction results.

The CORAL model for domain adaptive transfer learning from the perspective of feature transformations does not involve the selection of kernel functions, so this paper needs to combine the LLE dimensionality reduction algorithm to reduce the dimension of the source domain film sample features and uses the stacking ensemble learning model for regression to study the CORAL model. MAE and MSE of the film ratings from target domain are shown in Fig. 4.

When studying the CORAL model, PCA and ISOMAP are used for evaluating the LLE dimension reduction algorithm, KNN and XGBoost are selected for the stacking ensemble learning model. It can be seen from Fig. 4 that the CORAL model using LLE to reduce the dimension of the source domain film data and using the stacking ensemble learning model for regression has achieved the best prediction outcome, with an MAE of 1.14 and an MSE of 2.05. It is also found that when using KNN for regression, the prediction outcome for film rating after dimension reduction of the source domain film samples through LLE is poor, which is due to the insufficient fitting degree of KNN. The CORAL model studied in this paper has an average relative decrease of MAE by about 13.3% and MSE by about 13.5% compared with the comparative objects. Overall, the CORAL model by using LLE to reduce the dimension of the film sample data in source domain and using the stacking ensemble learning model for regression can also effectively predict the film ratings in target domain.



(a) The MAE of CORAL model prediction results

(b) The MSE of CORAL model prediction results

Fig. 4 The MAE and MSE of CORAL model prediction results

5 Conclusions

Focused on film rating predictions, this paper constructs new features such as composite type influence index, director level and actor level by appropriate data analysis methods and clustering algorithms on the IMDB film dataset. In order to solve problems that the feature dimensions of the source domain and the target domain film data are inconsistent, and the target domain film samples do not have film ratings, which leads to the inability to train a traditional machine learning model, this paper studies the TCA model with the PCA dimensionality reduction, linear kernel for mapping and stacking ensemble learning model for regression. The CORAL model with the LLE dimensionality reduction and stacking ensemble learning model for regression is also studied in this paper. Results show that when the domain adaptive transfer learning model is used with specific dimension reduction algorithms, there are certain advantages in solving the problem of film rating prediction without sample labels in the target domain and better prediction results can be achieved.

References

[1] LIU D, FANG J X. Key technologies and development strategies of film big data[J]. Contemporary Cinema, 2015(3): 117-121.

[2] YANG Q, ZHANG Y, DAI W Y, et al. Transfer Learning[M]. Beijing: China Machine Press, 2020: 6-11.

[3] PAN S J, YANG Q. Asurvey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010,22(10): 1345-1359.

[4] GAO S, XU Q Z. Overview of the application of transfer learning methods in the field of medical images[J]. Computer Engineering and Applications, 2021,57(24): 12.

[5] LI J J, MENG L C, ZHANG K, et al. Overview of domain adaptive research [J]. Computer Engineering, 2021,47(6): 1-13.

[6] CHENG J N, YU Z X, CHEN L, et al. Multi-source multi-label pedestrian attribute recognition based on domain adaptive[J]. Journal of Computer Applications, 2022,42(8): 2401-2406. (In Chinese)

[7] NOH K J, CHOI J, HONG J S, et al. Finger-vein recognition using heterogeneous databases by domain adaption based on a cycle-consistent adversarial network[J]. Sensors, 2021, 21(2): 524.

[8] ZHOU Z H. Machine learning[M]. Beijing: Tsinghua University Press, 2016: 75-78. (In Chinese)

[9] AURÉLIEN G. Hands-on machine learning with Scikit-Learn and Tensorflow[M]. Beijing: China Machine Press, 2018: 173-174.

[10] PAN S J, TSANG IVOR W, KWOK JAMES T, et al. Domain adaptation via transfer component analysis[J]. IEEE Transactions on Neural Networks, 2011,22(2): 199-210.

[11] WANG J D, CHEN Y Q. Introduction to transfor learning [M]. Beijing: Publishing House of Electronics Industry, 2021: 72-74. (In Chinese)

[12] HE L, CAI Y C, YANG Z. Summarization of high dimensional data clustering methods[J]. Application Research of Computers, 2010, 27(1): 23-26,31.

[13] SUN B, FENG J, SAENKO K. Return of frustratingly easy domain adaptation[C]//The 13th AAAI Conference on Artificial Intelligence. Phoenix: AAAI, 2016: 2058-2065.

[14] QIU J R, LUO H. Improved local linear embedding algorithm and application[J]. Computer Engineering and Applications,2020, 56(3): 176-179.

[15] ZHOU Z H. Basis and algorithm of ensemble learning [M]. Beijing: Publishing House of Electronics Industry, 2020: 14-16. (In Chinese)

[16] TANG K, QIN M, ZHAO X, et al. Prediction of gaseous nitrite based on stacking ensemble learning model[J]. China Environmental Science, 2020,40(2): 582-590. (In Chinese)

SHU Zhan, born in 1997. He is a graduate student in School of Information Science and Engineering, Shenyang University of Technology. He received his B. E. degree from Shenyang University of Technology in 2019. His research interests include machine learning and intelligent software.