Data Similarity Estimation for AI Model Reuse

Ling ZHOU † Tsuyoshi TANAKA † and Daisuke TASHIRO †

† R&D Group, Hitachi, Ltd. 1-280 Higashi-koigakubo, Kokubunji-shi, Tokyo, 185-8601 Japan

E-mail: † {ling.zhou.wm, tsuyoshi.tanaka.vz, daisuke.tashiro.em}@hitachi.com

Abstract In business expansion, two main problems exist in utilizing traditional machine learning techniques. (1) It is of high cost to operate and maintain a model for each customer. (2) It's difficult to learn a good model if training data is not enough. Model reuse is one solution. It reuses a model from previous (source) tasks to predict a new (target) task, where data similarity estimation is an issue in order to select the suitable model for reuse. Existing methods for image data can't be utilized on non-image data. In this paper, we propose to estimate data similarity by summarizing each feature's distribution distance with feature influence attached as weight. Experiment results on open non-image dataset verify that our proposed method can estimate data similarity well and the source model with the best performance on a target task can be selected for reuse successfully.

Keyword Transfer Learning, AI Model Reuse, Non-image Data, Data Similarity Estimation

1. Introduction

In business expansion, two main problems exist when traditional machine learning techniques are utilized to learn an AI model (hereafter 'model') for each customer (task) from its training data. (1) It is time-consuming and troublesome to operate and maintain a model for each customer. (2) It's difficult to learn a good model for the customer when the training data is not enough.

In order to solve them, model reuse is one approach. Model reuse is a kind of transfer learning [1], which has emerged as a new learning framework. Traditional machine learning techniques try to learn each task from scratch, while transfer learning techniques try to transfer the knowledge from some previous (source) tasks to a new (target) task. The knowledge can be model, data or exacted features. When a model is transferred from source to target, a source model will be reused to predict for the target task.

Data similarity estimation is one issue of model reuse. Data similarity estimation is utilized to select a suitable source model for reuse from a group of existing models by estimating the similarity between source data and target data. Related methods have been proposed in [2][3][4][5][6] for image data. In existing methods, each image (represented by pixels) is treated as one unit; distances between images are utilized for data similarity estimation. However, the structures of image data and non-image data are different. It's difficult to calculate the distance between data records directly because their feature characteristics and scales are different. So, existing methods for image data are not applicable to non-image data.

In this paper, we propose to estimate the data similarity between source data and target data by calculating each feature's distribution distance separately and summarizing them with feature influence considered. How to calculate feature distribution distance is also studied. Experiment results on UCI default of credit card clients dataset verify that our proposed method can estimate data similarity well and the source model with the best performance on a target task can be selected for reuse successfully.

2. Related work

Model reuse is a kind of transfer learning. It transfers a model from previous (source) tasks to predict a new (target) task. Take credit scoring service as an example. In the credit scoring system, some credit scoring models (e.g. Model A, Model B, Model C) have been trained previously for banks (e.g. Bank A, Bank B, Bank C) by using their training dataset (e.g. TrainData A, TrainData B, TrainData C). These models are treated as source models. When we want to provide the credit scoring service to a new bank (Bank T), we can reuse the source models instead of training a new model for Bank T. As a result, it saves our cost to operate and maintain a model for Bank T; and we can still provide a good service to Bank T when the training data in Bank T is not enough.

The process of model reuse is shown in Figure 1. First,

select a suitable source model from existing models as Model S according to the data similarity between source TrainData and target TrainData. Then, take measures to improve Model S's accuracy on target test data (TestData T). For example, we can fine-tune Model S by training Model S further using the training dataset of Bank T (TrainData T) to generate a new model Model T; then use Model T to predict TestData T. This paper focuses on the first step of model selection based on data similarity.

Data similarity estimation is utilized to select a suitable source model for reuse from a group of existing models by estimating the similarity between source data (e.g. TrainData A/B/C) and target data (e.g. TrainData T). Related methods have been proposed in [2][3][4][5][6] for image data. In existing methods, each image (represented by pixels) is treated as one unit; distances between images are utilized for data similarity estimation and domain adaptation. However, the structures of image data and non-image data are different. Take table data as an example. Each data record includes both numerical features and categorical features; and their scales are different. The features should be treated differently. It's difficult to calculate the distance between data records directly. So, existing methods of data similarity estimation are not applicable to non-image data. In this paper, we study related method for non-image data.

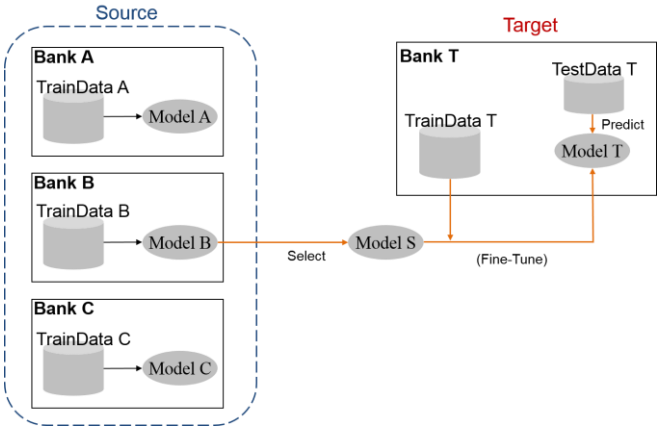

Figure 1. Process of model reuse

3. Proposed method

3.1 Overview of proposed method

Since data is represented by features, we think data similarity can be estimated by feature similarity. According to [13][14], each feature's influence on prediction is different. Some features are important, while others are unimportant. The similarity of important features contributes more to data similarity. So, feature influence should be considered when utilizing feature similarity to estimate data similarity. Our proposal is shown in Figure 2. We propose to calculate each feature's distribution distance separately to estimate feature similarity and summarize them with feature influence attached as weights. f_1, f_2, …, f_n are the features in target dataset TrainData T and source dataset TrainData S (S is A/B/C); $d_i^{T,S}$ is f_i's distribution distance between TrainData T and TrainData S. I_i^T is f_i's influence (weight) on task T prediction; it can be calculated in advance utilizing model information or eXplainable Artificial Intelligence (XAI) technology. $Distance_{T,S}$ is the dataset distance between TrainData T and TrainData S. It is used to represent their data similarity. The smaller $Distance_{T,S}$ is, the more similar TrainData T is to TrainData S. $Distance_{T,S}$ is calculated in the following equation.

$$Distance_{T,S} = I_1^T d_1^{T,S} + I_2^T d_2^{T,S} + \cdots + I_n^T d_n^{T,S} \quad (1)$$

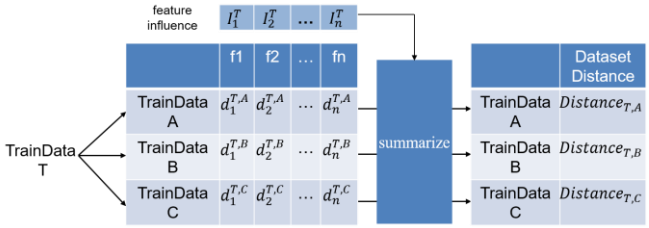

Figure 2. Proposed method of data similarity estimation

3.2 Feature distribution distance calculation

(1) Numerical feature

The scales of numerical features are different. The distribution distance of features with larger scales will be larger. In order to summarize them fairly for data similarity estimation, normalization is required before the calculation of distribution distance.

Numerical feature distribution distance can be calculated by popular measures of probability distribution distance [7] such as Kullback-Leibler (KL) Distance [8], Hellinger Distance [9][10] and Earth Mover's Distance (EMD) [11]. Compared to other methods, EMD is efficient because we don't need to estimate the probability density function of features in advance. So, we decided to utilize EMD, the minimum cost of tuning one probability distribution into the other, for numerical feature distribution distance calculation. Its comparison with other methods is a remaining issue that needs further study.

(2) Categorical feature

For categorical features, we represent them by the occurrence ratio of their candidate values. An example is as follows. In Bank A, there are 100 loan clients. Among them, 60 clients are male and 40 are female. So, the feature of 'gender' in Bank A can be represented by (0.6, 0.4). In Bank B, there are 200 loan clients. Among them, 100 clients are male and 100 are female. So, the feature of 'gender' in Bank B can be represented by (0.5, 0.5). Then, categorical feature distribution distance can be estimated by the Euclidean Distance (ED) between their occurrence ratio representations.

Since the distribution distances of numerical features and categorical features are calculated in different ways, they should be summarized separately when estimating the data similarity between TrainData T and TrainData S.

4. Evaluation

4.1 Dataset

The dataset we used for evaluation is default of credit card clients dataset [12]. The credit card issuer has gathered information on 30000 customers. The dataset contains information on 24 variables, including demographic factors, credit data, history of payment, and bill statements of credit card customers from April 2005 to September 2005, as well as information on the outcome: did the customer default or not.

4.2 Evaluation method

Before the evaluation of our proposed data similarity estimation method, we generated some artificial data to verify the applicability of EMD (for numerical features) and ED (for categorical features) on feature distribution distance calculation.

In order to evaluate the proposed data similarity estimation method, we separated default of credit card clients dataset into four groups according to the age of credit card client. Group 1 are the information on 9618 customers whose age is smaller than 30; Group 2 are the information on 10284 customers whose age is from 30 to 39; Group 3 are the information on 7418 customers whose age is from 39 to 50; and Group 4 are the information on 2680 customers whose age is larger than 50. For each group, we randomly selected 80% of its customer information as training data (TrainData) and used the remaining 20% as test data (TestData). Then we repeatedly used three of them as source and the remaining one as target to evaluate the performance of our proposed method for cross validation.

4.3 Evaluation of feature distribution distance calculation

4.3.1 Artificial data generation

Since we don't know the fact of feature similarity, we generated some artificial data, whose relative feature similarities are known, for verification.

We used part of default of credit card clients dataset as dataset A by picking up six variables as features. Dataset A includes three numerical features (f1 'LIMIT_BAL', f2 'BILL_AMT1', f3 'BILL_AMT2') and three categorical features (f4 'SEX', f5 'EDUCATION', f6 'MARRIAGE'). Next, we made a copy of dataset A to generate dataset A_artificial. Then we changed the values of features in dataset A_artificial and gradually increased the number of records for change. As shown in Figure 3, for numerical values, they are changed by multiplying them by 1.5; for categorical values, they are changed by adding them by 1. For example, for f1, no records are changed (the values of f1 in A and A_artificial are the same); for f3, 10000 records are changed. So, the feature distribution distances should also increase gradually. If the calculated feature distribution distances match with this trend, it will verify the effectiveness of EMD (for numerical features) and ED (for categorical features) on feature distribution distance calculation.

A	f1	f2	f3
A_artificial	f1*1.5	f2*1.5	f3*1.5
change_num	0	5000	10000

(a) Artificial data generation on numerical features

A	f4	f5	f6
A_artificial	f4+1	f5+1	f6+1
change_num	0	5000	10000

(b) Artificial data generation on categorical features

Figure 3. Artificial data generation

4.3.2 Experiment results

(1) Numerical features

The experiment results on numerical features are shown in Figure 4. The EMD increases gradually from f1 to f3. Moreover, we can see that the change trend of EMD matches with that of histograms. So, EMD is applicable to estimate the distribution distance of numerical features.

(2) Categorical features

The experiment results on categorical features are shown in Figure 5. The ED increases gradually from f4 to f6. Moreover, we can see that the change trend of ED matches

with that of histograms. So, our proposed method utilizing ED is applicable to estimate the distribution distance of categorical features.

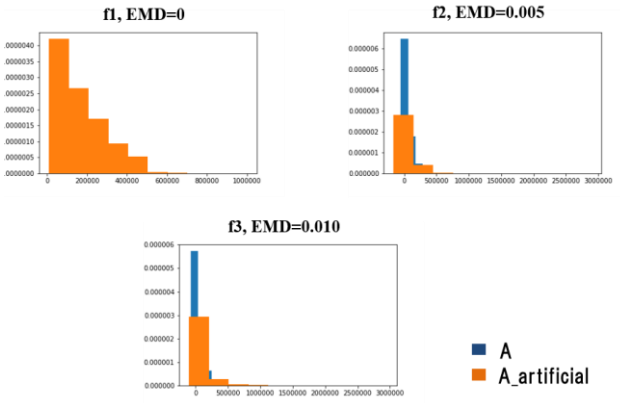

Figure 4. Feature distribution distances of numerical features

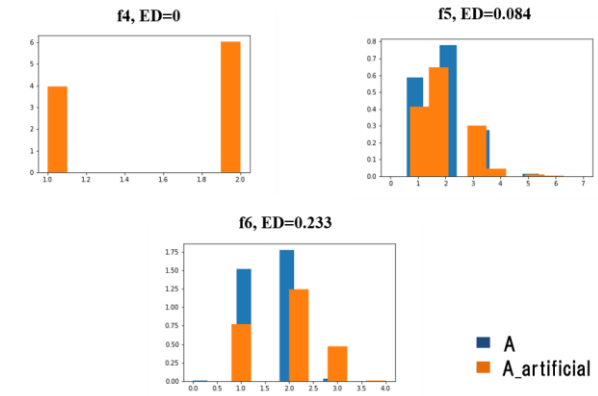

Figure 5. Feature distribution distances of categorical features

4.4 Evaluation of data similarity estimation

4.4.1 Experiment steps

We did experiments in the following steps to evaluate the effect of our proposed data similarity estimation method.

(Step 1) Select one group as the target (T) and use the remaining three groups as source (S1, S2, S3).

(Step 2) Train a LightGBM [13] model (Model S1, Model S2 and Model S3) for source S1, S2 and S3 separately by using their training dataset (TrainData S1, TrainData S2 and TrainData S3).

(Step 3) Calculate the feature influence (weight) for Target T (here, LightGBM.feature_importance() is utilized; other possible approaches include SHAP [14] and weights of Logistic Regression etc.); and perform normalization.

(Step 4) Repeat the following steps (4-1) (4-2) and (4-3) when S is S1, S2, S3.

(4-1) For each feature, calculate its distribution distance between TrainData T and TrainData S.

(4-2) Calculate $Distance_{T,S}$ by using Equation (1) to summarize the distribution distances of numerical features and categorical features separately.

(4-3) Use Model S to predict TestData T and calculate its accuracy (AUC(*100)) on TestData T.

4.4.2 Experiment results

(1) Group 4 is target T; Group 1, Group 2 and Group 3 are source S1, S2, and S3 correspondingly.

The experiment results are shown in Table 1. $Distance_{T,S3}$ on both numerical (num) features and categorical (cat) features is the smallest and Model S3's AUC on TestData T (75.79) is the highest, so it is verified that our proposed data similarity estimation method works well and it can select the suitable source model for reuse successfully.

We also trained a LightGBM model (Model T) by using TrainData T and evaluated its performance on TestData T as baseline. AUC of Model T on TestData T is 74.45, which is smaller than that of selected source model. So model reuse by our proposed method can effectively ensure good performance on target T even when its training data is not enough.

Table 1. Experiment results when Group 4 is target T

		S1	S2	S3
Dataset Distance to T	num	0.0145	0.0088	0.0042
	cat	0.1574	0.0954	0.0571
AUC(*100) on T		73.51	74.77	75.79

(2) Group 1 is target T; Group 2, Group 3 and Group 4 are source S1, S2, and S3 correspondingly.

The experiment results are shown in Table 2. For numerical (num) features, $Distance_{T,S2}$ is the smallest; for categorical (cat) features, $Distance_{T,S1}$ is the smallest. In this case, Model S1 and Model S2 are both candidate models for reuse. We applied both of them to predict TestData T and used the average of their predictions as final prediction results; and the final AUC reached to 78.89, higher than both Model S1's AUC (78.59) and Model S2's AUC (78.66). So, it is verified that our proposed data similarity estimation method can select the candidate models for reuse well.

We also trained a LightGBM model (Model T) by using TrainData T and evaluated its performance on TestData T as baseline. AUC of Model T on TestData T is 79.40. The

selected models' accuracy is near to the baseline; and by model reuse, the cost of model operation and maintenance is reduced.

(3) Group 2 is target T; Group 1, Group 3 and Group 4 are source S1, S2, and S3 correspondingly.

The experiment results are shown in Table 3. $Distance_{T,S2}$ on both numerical (num) features and categorical (cat) features is the smallest and Model S2's AUC on TestData T (77.60) is the highest. Moreover, it is also higher than the baseline AUC of target T (77.39).

(4) Group 3 is target T; Group 1, Group 2 and Group 4 are source S1, S2, and S3 correspondingly.

The experiment results are shown in Table 4. $Distance_{T,S2}$ on both numerical (num) features and categorical (cat) features is the smallest and Model S2's AUC on TestData T (79.38) is the highest. Moreover, it is also higher than the baseline AUC of target T (78.29).

Table 2. Experiment results when Group 1 is target T

		S1	S2	S3
Dataset Distance to T	num	0.0161	0.0111	0.0127
	cat	0.1263	0.1636	0.1661
AUC(*100) on T		78.59	78.66	77.50

Table 3. Experiment results when Group 2 is target T

		S1	S2	S3
Dataset Distance to T	num	0.0212	0.0033	0.0087
	cat	0.1491	0.0824	0.1359
AUC(*100) on T		76.83	77.60	74.33

Table 4. Experiment results when Group 3 is target T

		S1	S2	S3
Dataset Distance to T	num	0.0135	0.0031	0.0040
	cat	0.1415	0.0422	0.0574
AUC(*100) on T		78.18	79.38	76.22

4.5 Discussion

Experiment results on artificial data show that Earth Mover's Distance can measure the distribution distance of numerical features well and Euclidean Distance of occurrence ratio representations can measure the distribution distance of categorical features well. However, its comparison with other feature distribution calculation methods is not clear and needs more experiments.

For data similarity estimation, our proposed method summarizes feature distribution distances with feature influence as weights. Experiment results on default of credit card clients dataset show that it estimates the similarity between source data and target data well; and it can select the source model with the best performance on a target task for reuse successfully. In section 4.4.2(1)(3)(4), the ranks of numerical Dataset Distance and categorical Dataset Distance are the same; so the source model which is trained from the source dataset, whose numerical and categorical dataset distance to the target dataset is the smallest, is selected for reuse and its performance on the target task is the best. While, in section 4.4.2(2), the ranks of numerical Dataset Distance and categorical Dataset Distance are different. In this case, we reuse both candidate models to predict for the target task, use the average of their predictions as final prediction results, and get the highest accuracy. Here, when more than one model is selected, how to merge them still needs further study.

5. Conclusion and future work

We have proposed a method of data similarity estimation for non-image data by summarizing its feature distribution distances in order to realize AI model reuse; and evaluated it on an open dataset. Experiment results show that the data similarity estimation method can estimate the similarity between source data and target data well; it can select the source model with the best performance on a target task for reuse successfully. Moreover, we have studied the methods to calculate the distribution distance for numerical features and categorical features, which are required by data similarity estimation. Evaluation on artificial data show that Earth Mover's Distance can measure the distribution distance of numerical features well and Euclidean Distance of occurrence ratio representations can measure the distribution distance of categorical features well. Once calculated, they can be summarized to estimate data similarity.

In our experiments, we used feature influence as weights and attached them to feature distribution distances for data similarity estimation. In the future, we will study the necessity of using weights for summarization and other weights determination methods. Moreover, since the feature distribution distances of numerical features and categorical features are calculated in different ways, their distances should be summarized separately. When the ranks of numerical Dataset Distance and categorical Dataset Distance are the same, it is easy to select the source data with the highest similarity. When the ranks of numerical Dataset Distance and categorical Dataset Distance are different, studying how to merge them will be our future work.

References

[1] S. Pan and Q. Yang, “A Survey on Transfer Learning”, IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345-1359, Oct., 2010.

[2] Y. Cui, Y. Song, C. Sun, A. Howard and S. Belongie, “Large Scale Fine-Grained Categorization and Domain-Specific Transfer Learning”, 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, Jun. 18-22, 2018, arXiv:1806.06193 [cs.CV].

[3] J. Ngiam, D. Peng, V. Vasudevan, S. Kornblith, Q. V. Le and R. Pang, “Domain Adaptive Transfer Learning with Specialist Models”, arXiv:1811.07056 [cs.CV], Dec. 11, 2018.

[4] W. Ge and Y. Yu, “Borrowing Treasures from the Wealthy: Deep Transfer Learning through Selective Joint Fine-Tuning”, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, Jul. 21-26, 2017, arXiv:1702.08690 [cs.CV].

[5] Z. Cao, M. Long, J. Wang and M. I. Jordan, “Partial Transfer Learning with Selective Adversarial Networks”, 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, Jun. 18-22, 2018, arXiv:1707.07901 [cs.LG].

[6] Z. Ding, M. Shao and Y. Fu, “Incomplete Multisource Transfer Learning”, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, vol. 29, no. 2, Feb., 2018.

[7] M. Sugiyama, “Distance Approximation between Probability Distributions: Recent Advances in Machine Learning”, Transactions of the Japan Society for Industrial and Applied Mathematics, vol.23, no.3, pp.439-452, 2013.

[8] S. Kullback and R. A. Leibler, “On Information and Sufficiency”, The Annals of Mathematical Statistics, vol. 22, no. 1, pp. 79-86, Mar., 1951.

[9] T. Kailath, “The Divergence and Bhattacharyya Distance Measures in Signal Selection”, IEEE Transactions on Communication Technology, vol. 15, no. 1, pp. 52-60, Feb., 1967.

[10] A. Shemyakin, “Hellinger Distance and Non-informative Priors”, 2014 International Society for Bayesian Analysis, vol. 9, no. 4, pp. 923-938, 2014.

[11] E. Levina and P. Bickel, “The Earth Mover's Distance is the Mallows Distance: Some Insights from Statistics”, Proceedings of Eighth IEEE International Conference on Computer Vision (ICCV 2001), pp. 251-256, Vancouver, BC, Canada, Jul. 7-14, 2001.

[12] I.C. Yeh and C.H. Lien, “The Comparisons of Data Mining Techniques for the Predictive Accuracy of Probability of Default of Credit Card Clients”, Expert Systems with Applications, vol. 36, no. 2, pp. 2473-2480, Mar., 2009.

[13] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T.Y. Liu, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree”, 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, Dec. 4-9, 2017.

[14] S. M. Lundberg, S. Lee, “A Unified Approach to Interpreting Model Predictions”, 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, Dec. 4-9, 2017.