

PREDICTION MODEL TRAINING METHOD AND APPARATUS

TECHNICAL FIELD

[0001] One or more embodiments of this specification relate to the field of machine learning technologies, and in particular, to a prediction model training method and apparatus.

5

BACKGROUND

[0002] With the development of artificial intelligence (AI), deep neural networks (DNN) have gradually been applied to a plurality of fields such as risk assessment, speech recognition, face recognition, and natural language processing. A DNN network structure in different application scenarios is relatively fixed. To implement better model performance, more training data is
10 required.

[0003] In many fields such as medical treatment and finance, different enterprises or institutions have different data samples. Once the data is jointly trained, model precision is greatly improved, thereby effectively improving quality of service and improving user satisfaction. However, the original training data includes a large amount of user privacy and a large quantity of
15 business secrets, and any information leakage results in irreparable negative consequences. Therefore, while multi-party joint training is used to resolve a problem of data isolation, enhancing data privacy protection and fully mining data value have become a focus of research in recent years.

[0004] If training members who participate in joint training are located in different industries
20 and have a same user group, for example, between a payment platform and a social platform, the payment platform mainly holds payment data and the social platform mainly holds entertainment and social data, so that feature space of data collected by the two platforms is different and sample space is the same, which can be referred to as a vertical data segmentation scenario.

[0005] However, currently, a joint training manner applied to a vertical data segmentation
25 scenario is limited, and it is difficult to satisfy a higher requirement in actual application. Therefore, a joint training solution is required to better satisfy a requirement in actual application, for example,

improve a use effect of a model.

SUMMARY

[0006] Embodiments of this specification describe a prediction model training method and apparatus, to improve a use effect of a model.

5 **[0007]** According to a first aspect, a prediction model training method is provided, applied to a first party in a plurality of parties, including:

performing joint training with another party in the plurality of parties based on a first original feature and a sample label of a common sample in the plurality of parties, so that the plurality of parties correspondingly obtain a plurality of trained feature extraction models, and the
10 first party further obtains a trained fused feature processing model, where the another party holds another original feature of the common sample; obtaining a target fused feature of the common sample, where the target fused feature is obtained by performing fusion processing on a plurality of extracted features output for the common sample by the plurality of trained feature extraction models; processing the first original feature by using a target model to obtain a target extracted
15 feature; processing the target extracted feature by using the trained fused feature processing model to obtain a target prediction result; and training the target model with a goal of reducing a first difference between the target prediction result and the corresponding sample label and reducing a second difference between the target extracted feature and the target fused feature, where the target model is used to construct a trained prediction model with the fused feature processing model.

20 **[0008]** In an embodiment, the joint training includes: processing the first original feature by using a first feature processing model to obtain a first extracted feature; receiving a second extracted feature from any second party in the another party, where the second extracted feature is obtained by processing a second original feature of the common sample by using a second feature processing model; processing a first fused feature by using the fused feature processing model to
25 obtain a first prediction result, where the first fused feature is obtained by performing the fusion processing on the first extracted feature and each received second extracted feature; and training the fused feature processing model based on the first prediction result and the sample label, and calculating a backhaul gradient propagated to the plurality of feature extraction models, so that the plurality of parties each train the locally deployed feature extraction model by using a target
30 gradient determined based on the backhaul gradient.

[0009] In an embodiment, the fusion processing is concatenation, adding, or averaging.

[0010] In an embodiment, the training the target model with a goal of reducing a first difference between the target prediction result and the corresponding sample label and reducing a second difference between the target extracted feature and the target fused feature includes: determining a first training loss, where a calculation formula for the first training loss includes performing weighted summation on the first difference and the second difference, and the first training loss is positively correlated with both the first difference and the second difference, and a weight corresponding to the first difference is greater than a weight corresponding to the second difference; and updating a model parameter in the target model by using the first training loss.

[0011] In an embodiment, the method further includes: skipping training the fused feature processing model while training the target model.

[0012] In an embodiment, after the training the target model, the method further includes: fine-tuning the target model by using a sample feature and a sample label of a unique sample in the first party; and constructing a fine-tuned target model and the fused feature processing model into the trained prediction model.

[0013] In a specific embodiment, the fine-tuning the target model by using a sample feature and a sample label of a unique sample in the first party includes: determining, by using the first original feature of the common sample and the target model, feature extraction templates corresponding to label categories; processing the sample feature of the unique sample by using the target model to obtain a third extracted feature; processing the third extracted feature by using the fused feature processing model to obtain a third prediction result; and fine-tuning the target model with a goal of reducing a third difference between the third prediction result and the corresponding sample label and reducing a fourth difference between the third extracted feature and a feature extraction template corresponding to a label category of the unique sample.

[0014] Further, in an example, the determining, by using the first original feature of the common sample and the target model, feature extraction templates corresponding to label categories includes: processing a first original feature of each common sample by using the target model to obtain a corresponding fourth extracted feature; and performing, for the label categories, averaging on fourth extracted features corresponding to the label categories to obtain the corresponding feature extraction templates.

[0015] In another example, the fine-tuning the target model with a goal of reducing a third difference between the third prediction result and the corresponding sample label and reducing a

fourth difference between the third extracted feature and a feature extraction template corresponding to a label category of the unique sample includes: determining a second training loss, where a calculation formula for the second training loss includes performing weighted summation on the third difference and the fourth difference, and the second training loss is positively correlated with both the third difference and the fourth difference, and a weight corresponding to the third difference is greater than a weight corresponding to the fourth difference; and updating a model parameter in the target model by using the second training loss.

[0016] According to another aspect, in a specific embodiment, the method further includes: fine-tuning the fused feature processing model while fine-tuning the target model.

[0017] According to a second aspect, a prediction model training apparatus is provided and is integrated into a first party in a plurality of parties, including:

a joint training module, configured to perform joint training with another party in the plurality of parties based on a first original feature and a sample label of a common sample in the plurality of parties, so that the plurality of parties correspondingly obtain a plurality of trained feature extraction models, and the first party further obtains a trained fused feature processing model, where the another party holds another original feature of the common sample; a fused feature acquisition module, configured to obtain a target fused feature of the common sample, where the target fused feature is obtained by performing fusion processing on a plurality of extracted features output for the common sample by the plurality of trained feature extraction models; a feature extraction module, configured to process the first original feature by using a target model to obtain a target extracted feature; a prediction module, configured to process the target extracted feature by using the trained fused feature processing model to obtain a target prediction result; and a target training module, configured to train the target model with a goal of reducing a first difference between the target prediction result and the corresponding sample label and reducing a second difference between the target extracted feature and the target fused feature, where the target model is used to construct a trained prediction model with the fused feature processing model.

[0018] According to a third aspect, a computer-readable storage medium is provided. The computer-readable storage medium stores a computer program. When the computer program is executed in a computer, the computer is enabled to perform the method according to the first aspect.

[0019] According to a fourth aspect, a computing device is provided, and includes a memory and a processor. The memory stores executable code, and the processor executes the executable code to implement the method according to the first aspect.

[0020] In the method and apparatus provided in the embodiments of this specification, a label party in vertical split learning performs knowledge distillation (KD) on a common sample used in multi-party split learning to obtain an extended student model. Further, the student model can be fine-tuned by migrating embedding knowledge information of the common sample to a unique sample in the label party, thereby improving a use effect of a prediction model constructed based on the student model.

BRIEF DESCRIPTION OF DRAWINGS

[0021] To describe the technical solutions in the embodiments of the present invention more clearly, the following briefly describes the accompanying drawings required for describing the embodiments. Apparently, the accompanying drawings in the following description show merely some embodiments of the present disclosure, and a person of ordinary skill in the art can still derive other drawings from these accompanying drawings without creative efforts.

[0022] FIG. 1 is a schematic diagram of data segmentation according to an embodiment of this specification;

[0023] FIG. 2 is a schematic diagram of an implementation architecture for training a prediction model according to an embodiment of this specification;

[0024] FIG. 3 is a schematic flowchart of a prediction model training method according to an embodiment of this specification;

[0025] FIG. 4 is a schematic flowchart of a model fine-tuning method according to an embodiment of this specification; and

[0026] FIG. 5 is a schematic diagram of a prediction model training apparatus according to an embodiment of this specification.

DESCRIPTION OF EMBODIMENTS

[0027] The following describes the solutions provided in this specification with reference to the accompanying drawings.

[0028] Embodiments of this specification disclose a prediction model training solution. In this solution, a label holder (or referred to as a label party for short) in vertical split learning performs knowledge distillation on a common sample used in split learning, to obtain an extended student

model. Further, the student model can be fine-tuned by using a unique sample in the label party, thereby improving a use effect of a prediction model constructed based on the student model.

[0029] For ease of understanding, the following first describes data distribution in vertical split learning. A plurality of participants (or referred to as a plurality of data parties, a plurality of computing parties, or referred to as a plurality of parties for short) in vertical split learning respectively hold different feature parts of a sample. Some of samples held by the parties are same samples (or referred to as common samples), for example, a same user group, and other samples are unique samples of the parties. One of the plurality of parties also holds a sample label.

[0030] FIG. 1 is a schematic diagram of data segmentation according to an embodiment of this specification. A plurality of parties are shown as two parties: a data party A and a data party B. A label Y1 and a features X1 held by the data party A correspond to a same sample as a feature Z1 held by the data party B, and a sample corresponding to a label Y2 and a feature X2 are unique to the data party A.

[0031] It should be noted that there are a plurality of manners for determining a common sample. For example, a plurality of parties can jointly perform a private set intersection (PSI) protocol to obtain an intersection set of local sample sets of the parties as a common sample. Accordingly, a label holder can determine a sample other than the common sample in local samples as a unique sample.

[0032] The following uses the two parties shown in FIG. 1 as an example to describe different feature parts held by the parties and sample labels held by a label party in different service scenarios.

[0033] In a recommendation scenario, the data party A holds object feature data and a user behavior label of a recommended sample, and the data party B holds user feature data of the recommended sample. It should be noted that user-related data such as the user feature data mentioned in this specification is collected and used after a user confirms authorization.

[0034] In an embodiment, the data party A can be an e-commerce platform, a payment platform, a content platform, etc, and the object feature data owned by the data party A can be specific to a service object such as a commodity, a service, an official account, a subscription account, an application, content information, or an advertisement.

[0035] In an example, the object feature data is commodity features, such as a category, an origin, cost, selling price, sales volume, and target audience. In another example, the object feature

data is official account features, such as a category, a quantity of followers, a content post frequency, a quantity of likes, and a quantity of shares. In still another example, the object feature data is advertisement features, such as an advertisement carrier (including a picture, text, video, or audio), advertisement content, and spokesperson information.

5 **[0036]** In an embodiment, the user feature data includes a static attribute feature of a user, such as an occupation, interests and hobbies, permanent residence, and asset information. For example, these features can be extracted from information registered by the user. In another embodiment, the user data further includes a dynamic behavior feature of a user. In an example, the data party B is a social platform. In this case, the dynamic behavior feature can include social duration, a
10 social frequency, a quantity of friends, a quantity of personal posts, and a quantity of likes. In another example, the data party B is a payment platform. In this case, the dynamic behavior feature can include a payment frequency, a payment amount, and payment time.

[0037] The user behavior label indicates whether a user performs a predetermined behavior on a service object. It can be understood that the predetermined behavior is compatible with the
15 service object. In an embodiment, the service object is a commodity or a service, and in this case, the predetermined behavior can be "purchase" or "use". In another embodiment, the service object is a public account or a subscription account, and in this case, the predetermined behavior can be "follow" or "subscript". In still another embodiment, the service object is an application, and in this case, the predetermined behavior can be "download" or "register". In yet another embodiment,
20 the service object is an advertisement or content information, and in this case, the predetermined behavior can be "click".

[0038] Data distribution of the two parties in the recommendation scenario is described above.

[0039] In a risk control scenario, the two parties can respectively hold different behavior data of a same user, and the label party further holds a risk label of the user. In an embodiment, the two
25 parties include a bank and a payment platform. The bank holds a bank flow of a user, such as deposit and withdrawal records. The payment platform holds payment data of the user, such as a payment amount and payment time. In another embodiment, the two parties include a credit service platform and a credit institution. The credit service platform holds a record of a user using credit service data, for example, renting a book or a bicycle with no deposit. The credit institution holds
30 credit data of the user, for example, a quantity of credit cards held by the user and a default repayment record.

[0040] The risk label can be a risk category label, such as a low risk, a medium risk, and a high risk; or can be a risk score label, such as 30 points or 90 points on a 100-point scale.

[0041] Multi-party data distribution in a vertical split scenario is separately described above with reference to the recommendation scenario and the risk control scenario by using two parties as an examples. It can be understood that the plurality of parties are not limited to two parties, and the service scenario is not limited to the recommendation scenario and the risk control scenario. Exhaustive listing is not performed herein.

[0042] Next, an implementation architecture of a training solution according to an embodiment of this specification is described. FIG. 2 is a schematic diagram of an implementation architecture for training a prediction model according to an embodiment of this specification. Two parties in a plurality of parties are shown. A first feature extraction model and a fused feature processing model are locally deployed in a first party that holds a label in the two parties, and a second feature extraction model is locally deployed in a second party.

[0043] As shown in FIG. 2, in a forward propagation process during joint training (such as the above-mentioned vertical split learning), the first party processes features x_1 , x_2 , and x_3 of a common sample by using the first feature extraction model to obtain a first extracted feature H_A . The second party processes features z_4 , z_5 , and z_6 of the common sample by using the second feature extraction model to obtain a second extracted feature H_B . The first party processes a fused feature H_{AB} between the first extracted feature H_A and the second extracted feature H_B by using the fused feature processing model to obtain a prediction result for the common sample, and performs reverse propagation by using the prediction result and the sample label y , so that the first party and the second party respectively update the locally deployed models.

[0044] It should be understood that the above-mentioned extracted feature can also be referred to as an embedding feature or an embedding representation. In addition, "first" in the "first party", etc., "second" in the "second party", etc. "third" in another place of this specification, and similar terms all are used to distinguish a same thing, and do not have another limitation function such as sorting.

[0045] As shown in FIG. 2, after the joint training is completed, the first party performs knowledge distillation based on the fused feature H_{AB} between the extracted features output for the common sample by the trained first feature extraction model and second extraction model to obtain an extended student model. Therefore, the first party can directly process a local feature by

using the student model and the fused feature processing model without performing prediction in joint with another party, to achieve a prediction effect approximating that achieved in joint with another party.

[0046] With reference to FIG. 3 and more embodiments, the following describes specific implementation steps of the above-mentioned training solution. FIG. 3 is a schematic flowchart of a prediction model training method according to an embodiment of this specification. The method is applied to a first party in a plurality of parties. It should be understood that, each of the plurality of parties can be implemented as any apparatus, server, platform, device cluster, etc. that has a computing and processing capability. For example, the first party can be a payment platform.

[0047] As shown in FIG. 3, the method includes the following steps.

[0048] Step S310: Perform joint training with another party in the plurality of parties based on a first original feature and a sample label of a common sample, so that the plurality of parties correspondingly obtain a plurality of trained feature extraction models, and the first party further obtains a trained fused feature processing model, where the another party holds another original feature of the common sample. Step S320: Obtain a target fused feature of the common sample, where the target fused feature is obtained by performing fusion processing on a plurality of extracted features output for the common sample by the plurality of trained feature extraction models. Step S330: Process the first original feature by using a target model to obtain a target extracted feature. Step S340: Process the target extracted feature by using the trained fused feature processing model to obtain a target prediction result. Step S350: Train the target model with a goal of reducing a first difference between the target prediction result and the corresponding sample label and reducing a second difference between the target extracted feature and the target fused feature, where the target model is used to construct a trained prediction model with the fused feature processing model.

[0049] The above-mentioned steps are described below in detail.

[0050] First, in step S310, the first party performs the joint training with the another party in the plurality of parties based on the first original feature and the sample label of the common sample in the plurality of parties.

[0051] It can be understood that the first original feature refers to a local feature part held by the first party for the common sample (such as a same batch of users). There are usually a plurality of common samples, such as tens of thousands. There are one or more other parties, which

respectively hold different original features of the common sample. For the sample label of the common sample and different original features held by the plurality of parties, references can be made to the above-mentioned description of data distribution in a plurality of parties. Details are not described herein again.

5 **[0052]** Through the joint training (or vertical split learning), each of the plurality of parties completes training on the locally deployed feature extraction model, where the feature extraction model is used to perform feature extraction on an original feature; and the first party further completes training on the locally deployed fused feature processing model, where the fused feature processing model is used to process a fused feature between a plurality of extracted features output
10 from a same sample by the plurality of feature extraction models that are correspondingly deployed in the plurality of parties, to obtain a prediction result for the same sample.

[0053] The feature extraction models respectively deployed in the plurality of parties can be based on a same algorithm or different algorithms, and the algorithm can be specifically a DNN network, a convolutional neural network (CNN), a recurrent neural network (RNN), etc. An
15 implementation algorithm of the fusion processing model deployed in the first party can also be a DNN, a CNN, an RNN network, etc. Generally, the feature extraction model is a lightweight model, and can be, for example, implemented as a multilayer perceptron (MLP for short) that includes one or two hidden layers. The fusion processing model is relatively complex, and can be, for example, implemented as a DNN network with a relatively large quantity of network layers.

20 **[0054]** A manner of performing the joint training in this step is described below by using an example, and an existing manner can be selected based on a requirement in practice. This is not limited in this specification embodiment.

[0055] For example, the joint training process includes the following steps: The first party processes the first original feature by using a first feature processing model to obtain a first
25 extracted feature; the first party receives a second extracted feature from each party (or referred to as a second party) in the another party, where the second extracted feature is obtained by the second party by processing a second original feature of the common sample by using a second feature processing model; the first party processes a first fused feature by using the fused feature processing model to obtain a first prediction result, where the first fused feature is obtained by
30 performing the fusion processing on the first extracted feature and each received second extracted feature; and the first party trains the fused feature processing model based on the first prediction

result and the corresponding sample label, and calculates a backhaul gradient propagated to the plurality of feature extraction models, so that the plurality of parties each train the locally deployed feature extraction model by using a target gradient determined based on the backhaul gradient.

[0056] It can be understood that, there are a plurality of manners for the above-mentioned fusion processing, such as concatenation, summation, and averaging. Actually, the fusion processing may be more complex, such as weighted summation, where a weight may be set manually, a training parameter may be set, or an attention mechanism may be introduced to determine a weight. In addition, a manner in which the first party calculates the target gradient based on the backhaul gradient is compatible with a fusion processing manner. For example, the fusion processing manner is concatenation. In this case, the backhaul gradient can be directly split into a plurality of sub-vectors to be correspondingly provided to the plurality of parties as a plurality of target gradients. For another example, the fusion processing manner is averaging. In this case, the backhaul gradient can be directly provided to the plurality of parties as the target gradient.

[0057] Through the above, the first party can obtain a trained first feature extraction model and fused feature processing model by participating in the joint training, and the another party obtains a trained local feature extraction model by participating in the joint training.

[0058] In step S320, the target fused feature of the common sample is obtained, where the target fused feature is obtained by performing the fusion processing on the plurality of extracted features output for the common sample by the plurality of trained feature extraction models.

[0059] In an implementation, the first party can first obtain the plurality of extracted features. Specifically, the first party processes the first original feature by using the trained first feature extraction model to obtain a corresponding extracted feature, and the first party receives, from the second party in the plurality of parties, an extracted feature obtained by the second party by processing the second original feature of the common sample by using a trained second feature processing model. Therefore, the first party can obtain the plurality of extracted features of the common sample.

[0060] Further, the first party can perform the fusion processing on the plurality of extracted features to obtain the target fused feature. It should be noted that an implementation of the fusion processing is the same as a fusion processing manner used in the previous joint training. For details, references can be made to related descriptions in the above-mentioned embodiment. Details are

not described again.

[0061] In another implementation, when obtaining authorization from the another party, the first party can directly read a fused feature generated in a process of traversing the common sample for the last time in the joint training, and use the fused feature as the target fused feature.

5 **[0062]** Through the above, the first party can obtain the target fused feature of the common sample.

[0063] Next, the target fused feature is used as teacher knowledge in knowledge distillation to train the target model, or referred to as a student model, by using step S330, step S340, and step S350.

10 **[0064]** In step S330, the first original feature is processed by using the target model to obtain the target extracted feature.

[0065] It should be understood that, a function of the target model is also feature extraction (or feature embedding), and an output dimension of the target model (for example, a dimension of the target extracted feature) is the same as an input dimension of the fused feature processing model (for example, a dimension of the target fused feature). Therefore, the fused feature processing model can process output of the target model. For example, an implementation algorithm of the target model can be an MLP, a DNN, a CNN, an RNN, etc.

[0066] In this way, the first original feature can be input to the target model to obtain the target extracted feature. Therefore, in step S340, the target extracted feature is processed by using the trained fused feature processing model to obtain the target prediction result.

[0067] Then, in step S350, the target model is trained with the goal of reducing the first difference between the target prediction result and the corresponding sample label and reducing the second difference between the target extracted feature and the target fused feature.

[0068] Specifically, a first training loss is first determined based on the first difference and the second difference. It can be understood that the first training loss is positively correlated with both the first difference and the second difference.

[0069] In an embodiment, weighted summation is performed on the first difference and the second difference to obtain the first training loss. In a specific embodiment, a weight corresponding to the first difference is greater than a weight corresponding to the second difference.

30 For example, a calculation formula for the first training loss is constructed as follows:

$$L1=L0(Y1_pred,Y1)+u1*\|H_AB'-H_AB\| \quad (1)$$

[0070] In the above-mention formula (1), $L1$ represents the first training loss; $Y1_pred$ represents the target prediction result, and $Y1$ represents the corresponding sample label; $L0(Y1_pred, Y1)$ measures the first difference, which may be for example, a cross-entropy loss; H_AB' represents the target extracted feature, H_AB represents the target fused feature, the
5 operator $\| \|$ represents an L2 norm of a calculation vector, and $\|H_AB' - H_AB\|$ measures the second difference; and $U1$ represents the weight of the second difference, which is a hyperparameter whose value range is an interval $(0, 1)$. For example, $u1=0.01$ may be set.

[0071] It should be understood that, construction of the first training loss is described above by using an example. Actually, there are a plurality of other manners. For example, the
10 hyperparameter $u1$ is removed from the formula (1), or a cosine distance between H_AB' and H_AB is calculated to measure the second difference.

[0072] Further, a parameter in the target model can be adjusted by using the determined first training loss. It should be understood that, parameter adjustment, that is, model parameter adjustment and update, can be implemented by using a reverse propagation method. Details are
15 not described herein. In addition, it should be noted that in the process of training the target model in this step, a parameter in the fused feature processing model is frozen and fixed, and the fused feature processing model is not trained.

[0073] Through the above, a trained target model can be obtained.

[0074] Further, in an implementation, the trained target model and the fused feature processing
20 model can be directly constructed as the trained prediction model. Use of the prediction model includes the following: The first party inputs a local feature of a to-be-predicted sample (a feature item of the local feature is the same as a feature item of the first original feature) to the target model to obtain a corresponding extracted feature, and then inputs the extracted feature to the fused feature processing model to obtain a corresponding prediction result. It should be understood
25 that when using the prediction model, the first party does not need to use an original feature or extracted feature provided by the another party. In addition, through distillation learning, output of the target model approximates a fused feature between the plurality of parties, so that output of the fused feature processing model also approximates a prediction result obtained based on a multi-party fused feature, thereby greatly improving availability of the model, and further reducing a
30 calculation amount and a communication amount.

[0075] In another implementation, a unique sample in the first party is not used to perform

model training, and as a result, information about the sample is not learned by the model. This is especially wasteful when there is a relatively large quantity of unique samples. Therefore, it is further proposed in this embodiment of this specification that, the trained target model and/or fused feature processing model can be fine-tuned by using a unique sample, to further improve an overall model effect.

[0076] Specifically, after step S350 is performed, the method can further include step S360, that is, the target model is fine-tuned by using a sample feature and a sample label of a unique sample in the first party. Therefore, a fine-tuned target model and a fine-tuned fused feature processing model can be constructed as the trained prediction model.

[0077] In an embodiment, common samples can be further introduced into the fine tuning. Specifically, embedding knowledge information of different categories are obtained based on different classification labels of the common samples, and the embedding knowledge information is distilled into unique sample data, to effectively improving an overall model effect.

[0078] FIG. 4 is a schematic flowchart of a model fine-tuning method according to an embodiment of this specification. As shown in FIG. 4, the first party implements model fine-tuning by performing the following steps: S41, S42, S43, and S44.

[0079] Step S41: Determine, by using the first original feature of the common sample and the target model, feature extraction templates corresponding to label categories.

[0080] Specifically, the first original feature of each common sample is first processed by using the target model to obtain a corresponding fourth extracted feature. Then, the corresponding feature extraction templates are determined for the label categories based on fourth extracted features corresponding to the label categories.

[0081] Further, in an example, all fourth extracted features in the label category can be averaged, so that an obtained average feature is used as the feature extraction template in the label category. In another example, all fourth extracted features in the label category can be clustered, and an obtained cluster center or an average vector of a plurality of clusters is used as the feature extraction template in the label category.

[0082] In this way, the feature extraction templates corresponding to the label categories can be obtained based on the common sample. In another aspect, in step S42, the sample feature of the unique sample is processed by using the target model to obtain a third extracted feature, and in step S43, the third extracted feature is processed by using the fused feature processing model to

obtain a third prediction result.

[0083] Based on the feature extraction template determined by using the common sample and the obtained third extracted feature and third prediction result corresponding to the unique sample, step S44 is performed to fine-tune the target model with a goal of reducing a third difference between the third prediction result and the corresponding sample label and reducing a fourth difference between the third extracted feature and a feature extraction template corresponding to a label category of the unique sample.

[0084] Specifically, a second training loss is first determined based on the third difference and the fourth difference. It can be understood that the second training loss is positively correlated with both the third difference and the fourth difference.

[0085] In an embodiment, weighted summation is performed on the third difference and the fourth difference to obtain the second training loss. In a specific embodiment, a weight corresponding to the third difference is greater than a weight corresponding to the fourth difference. For example, a calculation formula for the second training loss is constructed as follows:

$$L2=L0(Y2_pred,Y2)+u2*\|Ei-ei\| \quad (2)$$

[0086] $L2$ represents the second training loss; $Y2_pred$ represents the target prediction result, and $Y2$ represents the corresponding sample label; $L0(Y2_pred,Y2)$ measures the third difference, which may be, for example, a cross-entropy loss; Ei represents a feature extraction template in an i^{th} label category, ei represents an extracted feature of a unique sample (whose label is the i^{th} label category), the operator $\| \|$ represents an L2 norm of a calculation vector, and $\|Ei-ei\|$ measures the fourth difference; and $U2$ represents the weight of the fourth difference, which is a hyperparameter whose value range is an interval $(0,1)$. For example, $u2=0.005$ may be set.

[0087] It should be understood that, construction of the second training loss is described above by using an example. Actually, there are a plurality of other manners. For example, the hyperparameter $u2$ is removed from the formula (2), or a cosine distance between Ei and ei is calculated to measure the fourth difference.

[0088] According to another aspect, in an embodiment, the fused feature processing model can be fine-tuned together.

[0089] A manner of fine-tuning the target model by distilling embedding knowledge information of common samples of different categories into unique sample data is described above with reference to FIG. 4. It should be understood that an execution sequence of the steps in FIG.

4 is not unique, and only needs to conform to data flow logic.

[0090] In another embodiment, the target model can be fine-tuned by using only a unique sample. Specifically, after the sample feature of the unique sample is processed by using the target model and the fused feature processing model to obtain the corresponding prediction result, a parameter in the target model is updated by using the prediction result and the corresponding sample label. For example, the target model can be fine-tuned by using only the item $L0(Y2_pred, Y2)$ in the formula (2).

[0091] In still another embodiment, sample features of the common sample and the unique sample can be first processed by using the target model and the fused feature processing model to obtain a corresponding prediction result, to fine-tune the target model and/or the fused feature processing model by using the prediction result and the corresponding sample label of the common sample, and the prediction result and the corresponding sample label of the unique sample.

[0092] Through the above, the unique sample in the first party can be used for model training, thereby improving an overall model effect.

[0093] Corresponding to the above-mentioned training method, an embodiment of this specification further discloses a training apparatus. FIG. 5 is a schematic diagram of a structure of a prediction model training apparatus according to an embodiment of this specification. The illustrated training apparatus 500 is integrated into a first party in a plurality of parties. As shown in FIG. 5, the training apparatus 500 includes the following modules:

a joint training module 510, configured to perform joint training with another party in the plurality of parties based on a first original feature and a sample label of a common sample in the plurality of parties, so that the plurality of parties correspondingly obtain a plurality of trained feature extraction models, and the first party further obtains a trained fused feature processing model, where the another party holds another original feature of the common sample; a fused feature acquisition module 520, configured to obtain a target fused feature of the common sample, where the target fused feature is obtained by performing fusion processing on a plurality of extracted features output for the common sample by the plurality of trained feature extraction models; a feature extraction module 530, configured to process the first original feature by using a target model to obtain a target extracted feature; a prediction module 540, configured to process the target extracted feature by using the trained fused feature processing model to obtain a target prediction result; and a target training module 550, configured to train the target model with a goal of reducing a first difference between the target prediction result and the corresponding sample label and reducing a second difference between the target extracted feature and the target fused

feature, where the target model is used to construct a trained prediction model with the fused feature processing model.

[0094] In an embodiment, the joint training module 510 is specifically configured to:

5 process the first original feature by using a first feature processing model to obtain a first extracted feature; receive a second extracted feature from any second party in the another party, where the second extracted feature is obtained by processing a second original feature of the common sample by using a second feature processing model; process a first fused feature by using the fused feature processing model to obtain a first prediction result, where the first fused feature is obtained by performing the fusion processing on the first extracted feature and each received
10 second extracted feature; and train the fused feature processing model based on the first prediction result and the sample label, and calculate a backhaul gradient propagated to the plurality of feature extraction models, so that the plurality of parties each train the locally deployed feature extraction model by using a target gradient determined based on the backhaul gradient.

[0095] In an embodiment, the fusion processing is concatenation, adding, or averaging.

15 **[0096]** In an embodiment, the target training module 550 is specifically configured to: determine a first training loss, where a calculation formula for the first training loss includes performing weighted summation on the first difference and the second difference, and the first training loss is positively correlated with both the first difference and the second difference, and a weight corresponding to the first difference is greater than a weight corresponding to the second
20 difference; and update a model parameter in the target model by using the first training loss.

[0097] In an embodiment, the target training module 550 is further configured to skip train the fused feature processing model while train the target model.

[0098] In an embodiment, the training apparatus 500 further includes: a fine-tuning module 560, configured to fine-tune the target model by using a sample feature and a sample label of a
25 unique sample in the first party; and a model construction module 570, configured to construct a fine-tuned target model and the fused feature processing model into the trained prediction model.

[0099] In a specific embodiment, the fine-tuning module 560 is specifically configured to: determine, by using the first original feature of the common sample and the target model, feature extraction templates corresponding to label categories; process the sample feature of the unique
30 sample by using the target model to obtain a third extracted feature; process the third extracted feature by using the fused feature processing model to obtain a third prediction result; and fine-tune the target model with a goal of reducing a third difference between the third prediction result

and the corresponding sample label and reducing a fourth difference between the third extracted feature and a feature extraction template corresponding to a label category of the unique sample.

[0100] Further, in an example, the fine-tuning module 560 is further configured to: process a first original feature of each common sample by using the target model to obtain a corresponding fourth extracted feature; and perform, for the label categories, averaging on fourth extracted features corresponding to the label categories to obtain the corresponding feature extraction templates.

[0101] In another example, the fine-tuning module 560 is further configured to: determine a second training loss, where a calculation formula for the second training loss includes performing weighted summation on the third difference and the fourth difference, and the second training loss is positively correlated with both the third difference and the fourth difference, and a weight corresponding to the third difference is greater than a weight corresponding to the fourth difference; and update a model parameter in the target model by using the second training loss.

[0102] In another specific embodiment, the fine-tuning module 560 is further configured to fine-tune the fused feature processing model while fine-tuning the target model.

[0103] An embodiment of another aspect further provides a computer-readable storage medium. The computer-readable storage medium stores a computer program, and when the computer program is executed in a computer, the computer is enabled to perform the method described with reference to FIG. 3 or FIG. 4.

[0104] An embodiment of still another aspect further provides a computing device, including a memory and a processor. The memory stores executable code, and the processor executes the executable code to implement the method described with reference to FIG. 3 or FIG. 4. A person skilled in the art should be aware that in the above-mentioned one or more examples, functions described in the present invention can be implemented by hardware, software, firmware, or any combination thereof. When the functions are implemented by using software, the functions can be stored in a computer-readable medium or transmitted as one or more instructions or code in the computer-readable medium.

[0105] In the above-mentioned specific implementations, the objectives, technical solutions, and beneficial effects of the present invention are further described in detail. It should be understood that the above-mentioned descriptions are merely specific implementations of the present invention, but are not intended to limit the protection scope of the present invention. Any

modification, equivalent replacement, improvement, etc. made based on the technical solutions of the present invention shall fall within the protection scope of the present invention.

CLAIMS

What is claimed is:

1. A prediction model training method, applied to a first party in a plurality of parties, comprising:

5 performing joint training with another party in the plurality of parties based on a first original feature and a sample label of a common sample in the plurality of parties, so that the plurality of parties correspondingly obtain a plurality of trained feature extraction models, and the first party further obtains a trained fused feature processing model, wherein the another party holds another original feature of the common sample;

10 obtaining a target fused feature of the common sample, wherein the target fused feature is obtained by performing fusion processing on a plurality of extracted features output for the common sample by the plurality of trained feature extraction models;

processing the first original feature by using a target model to obtain a target extracted feature;

15 processing the target extracted feature by using the trained fused feature processing model to obtain a target prediction result; and

training the target model with a goal of reducing a first difference between the target prediction result and the corresponding sample label and reducing a second difference between the target extracted feature and the target fused feature, wherein the target model is used to construct a trained prediction model with the fused feature processing model.

20 2. The method according to claim 1, wherein the joint training comprises:

processing the first original feature by using a first feature processing model to obtain a first extracted feature;

25 receiving a second extracted feature from any second party in the another party, wherein the second extracted feature is obtained by processing a second original feature of the common sample by using a second feature processing model;

processing a first fused feature by using the fused feature processing model to obtain a first prediction result, wherein the first fused feature is obtained by performing the fusion processing on the first extracted feature and each received second extracted feature; and

30 training the fused feature processing model based on the first prediction result and the sample label, and calculating a backhaul gradient propagated to the plurality of feature extraction models,

so that the plurality of parties each train the locally deployed feature extraction model by using a target gradient determined based on the backhaul gradient.

3. The method according to claim 1, wherein the fusion processing is concatenation, adding, or averaging.

5 4. The method according to claim 1, wherein the training the target model with a goal of reducing a first difference between the target prediction result and the corresponding sample label and reducing a second difference between the target extracted feature and the target fused feature comprises:

10 determining a first training loss, wherein a calculation formula for the first training loss comprises performing weighted summation on the first difference and the second difference, and the first training loss is positively correlated with both the first difference and the second difference, and a weight corresponding to the first difference is greater than a weight corresponding to the second difference; and

 updating a model parameter in the target model by using the first training loss.

15 5. The method according to claim 1, further comprising:

 skipping training the fused feature processing model while training the target model.

6. The method according to claim 1, wherein after the training the target model, the method further comprises:

20 fine-tuning the target model by using a sample feature and a sample label of a unique sample in the first party; and

 constructing a fine-tuned target model and the fused feature processing model into the trained prediction model.

7. The method according to claim 6, wherein the fine-tuning the target model by using a sample feature and a sample label of a unique sample in the first party comprises:

25 determining, by using the first original feature of the common sample and the target model, feature extraction templates corresponding to label categories;

 processing the sample feature of the unique sample by using the target model to obtain a third extracted feature;

30 processing the third extracted feature by using the fused feature processing model to obtain a third prediction result; and

 fine-tuning the target model with a goal of reducing a third difference between the third

prediction result and the corresponding sample label and reducing a fourth difference between the third extracted feature and a feature extraction template corresponding to a label category of the unique sample.

8. The method according to claim 7, wherein the determining, by using the first original
5 feature of the common sample and the target model, feature extraction templates corresponding to label categories comprises:

processing a first original feature of each common sample by using the target model to obtain a corresponding fourth extracted feature; and

performing, for the label categories, averaging on fourth extracted features corresponding to
10 the label categories to obtain the corresponding feature extraction templates.

9. The method according to claim 7, wherein the fine-tuning the target model with a goal of reducing a third difference between the third prediction result and the corresponding sample label and reducing a fourth difference between the third extracted feature and a feature extraction template corresponding to a label category of the unique sample comprises:

15 determining a second training loss, wherein a calculation formula for the second training loss comprises performing weighted summation on the third difference and the fourth difference, and the second training loss is positively correlated with both the third difference and the fourth difference, and a weight corresponding to the third difference is greater than a weight corresponding to the fourth difference; and

20 updating a model parameter in the target model by using the second training loss.

10. The method according to claim 6, further comprising:

fine-tuning the fused feature processing model while fine-tuning the target model.

11. A prediction model training apparatus, integrated into a first party in a plurality of parties, comprising:

25 a joint training module, configured to perform joint training with another party in the plurality of parties based on a first original feature and a sample label of a common sample in the plurality of parties, so that the plurality of parties correspondingly obtain a plurality of trained feature extraction models, and the first party further obtains a trained fused feature processing model, wherein the another party holds another original feature of the common sample;

30 a fused feature acquisition module, configured to obtain a target fused feature of the common sample, wherein the target fused feature is obtained by performing fusion processing on a plurality

of extracted features output for the common sample by the plurality of trained feature extraction models;

a feature extraction module, configured to process the first original feature by using a target model to obtain a target extracted feature;

5 a prediction module, configured to process the target extracted feature by using the trained fused feature processing model to obtain a target prediction result; and

a target training module, configured to train the target model with a goal of reducing a first difference between the target prediction result and the corresponding sample label and reducing a second difference between the target extracted feature and the target fused feature, wherein the
10 target model is used to construct a trained prediction model with the fused feature processing model.

12. A computer-readable storage medium, wherein the computer-readable storage medium stores a computer program, and when the computer program is executed in a computer, the computer is enabled to perform the method according to any one of claims 1 to 10.

15 13. A computing device, comprising a memory and a processor, wherein the memory stores executable code, and the processor executes the executable code to implement the method according to any one of claims 1 to 10.