BMC Biotechnology

RESEARCH

Open Access

Soft sensor modeling method for *Pichia pastoris* fermentation process based on substructure domain transfer learning

Bo Wang[1], Jun Wei[1*], Le Zhang[2], Hui Jiang[1], Cheng Jin[2] and Shaowen Huang[1]

Abstract

Background Aiming at the problem that traditional transfer methods are prone to lose data information in the overall domain-level transfer, and it is difficult to achieve the perfect match between source and target domains, thus reducing the accuracy of the soft sensor model.

Methods This paper proposes a soft sensor modeling method based on the transfer modeling framework of substructure domain. Firstly, the Gaussian mixture model clustering algorithm is used to extract local information, cluster the source and target domains into multiple substructure domains, and adaptively weight the substructure domains according to the distances between the sub-source domains and sub-target domains. Secondly, the optimal subspace domain adaptation method integrating multiple metrics is used to obtain the optimal projection matrices W_s and W_t that are coupled with each other, and the data of source and target domains are projected to the corresponding subspace to perform spatial alignment, so as to reduce the discrepancy between the sample data of different working conditions. Finally, based on the source and target domain data after substructure domain adaptation, the least squares support vector machine algorithm is used to establish the prediction model.

Results Taking *Pichia pastoris* fermentation to produce inulinase as an example, the simulation results verify that the root mean square error of the proposed soft sensor model in predicting *Pichia pastoris* concentration and inulinase concentration is reduced by 48.7% and 54.9%, respectively.

Conclusion The proposed soft sensor modeling method can accurately predict *Pichia pastoris* concentration and inulinase concentration online under different working conditions, and has higher prediction accuracy than the traditional soft sensor modeling method.

Keywords Substructure domain, Transfer learning, Soft sensor, *Pichia pastoris*

Background

As one of the most widely used exogenous protein expression systems [1, 2], *Pichia pastoris* (eukaryotic) expression system has achieved remarkable results in the fields of drug research and development, vaccine production, and industrial enzymes due to its simplicity of operation, high efficiency of expression, ease of cultivation, and the ability to post-transcriptional modifications of exogenous protein [3–5]. However, the process of protein production by *Pichia pastoris* induced fermentation is a highly nonlinear and strongly coupled dynamic process

*Correspondence:
Jun Wei
2222207065@stmail.ujs.edu.cn
[1] Key Laboratory of Agricultural Measurement and Control Technology and Equipment for Mechanical Industrial Facilities, School of Electrical and Information Engineering, Jiangsu University, Zhenjiang 212013, China
[2] Wuxi Key Laboratory of Intelligent Robot and Special Equipment Technology, Wuxi Taihu University, Wuxi 214064, China

with time-variability, strong coupling and uncertainty [6]. Key biological variables (such as *Pichia pastoris* concentration and inulinase concentration) that can directly reflect fermentation quality during the fermentation process cannot be measured online and in real time, and there is no accurate mechanism model available [7]. At present, it can only be estimated by offline and laboratory analysis, which not only causes the lag of information acquisition, and affects the operator's correct judgment and decision on the real-time reaction state, but also limits the implementation of optimal control strategy. Therefore, it is urgent to find a method to achieve the optimal estimation and prediction of key biological variables in the fermentation process of *Pichia pastoris*.

Soft sensor method is an effective way to address the problem of online measurement of key biological variables in biological fermentation process. Sun et al. [8] proposed a modeling method combining self-organizing feature mapping and least squares support vector machine to predict the fermentation effect of CTC. Experiments showed that the method could obtain more accurate predictions of fermentation effects. Wang et al. [9] applied relevance vector machine to the soft sensor modeling of penicillin fermentation process and achieved good results. Hua et al. [10] proposed a soft sensor model of penicillin fermentation process based on random forest and improved harris hawk optimized long short-term memory network to determine key biological variables in the fermentation process. The simulation results show that the established soft sensor model has high measurement accuracy and good measurement effect, and can meet the practical requirements of engineering. Dave et al. [11] used artificial neural networks and genetic algorithms to predict bioethanol production. Yamada et al. [12] used Gaussian mixture model to divide the datasets, genetic algorithm to select explanatory variables, and ultimately constructed online nonlinear adaptive soft sensor model for explanatory variables at each stage. The results show that the adaptive soft sensor model can accurately predict the value of the target variable in each process state. Although the soft sensor models constructed above can realize the online prediction of the key biological variables in the fermentation process, these modeling methods do not take into account the characteristics of multiple working conditions of the fermentation process, i.e., due to the different initial environmental parameters in the fermentation production process and the frequently switched parameters in the production process, there are large discrepancies between the fermentation data of different batches, and the data of fermentation process under different working conditions have drifted to a certain extent, and its distribution no longer obeys the assumption of independent and identical distribution, and it is difficult to collect labeled data for some special working conditions or potential working conditions, and when the distribution of the working conditions to be measured differs greatly from the distribution of the modeled data, the performance of the originally established soft sensor model will be significantly reduced, and the generalization capability will be limited or even the model will be invalidated, and the performance will be challenged considerably.

At present, although most soft sensor modeling algorithms considering multiple working conditions are relatively mature, they are still not rid of the assumption of independent and identical distribution in essence, and cannot break through the limitations of low prediction accuracy and poor generalization ability of the original model due to the discrepancy in data distribution of the working conditions to be measured under non-ideal conditions. The soft sensor method of multiple working condition process based on transfer learning solves the above problems. It relaxes the assumption that training data and test data need to follow independent and identical distribution, and quickly improves the accuracy of soft sensor model by transferring data information of known working conditions to the working condition to be measured with different data distribution and scarce labeled data, which is suitable for complex multiple working conditions. Chai et al. [13] proposed a deep probability transfer regression soft sensor framework, which reduced the discrepancy in data distribution between source domain and target domain and effectively reduced the impact of data loss on the performance of soft sensor models in industrial processes. Xie et al. [14] proposed an online transfer learning technology based on transfer slow feature analysis and variational Bayesian inference to solve the problem of measuring the water content of crude oil emulsion in steam-assisted gravity drainage technology. Ren et al. [15] proposed a soft sensor model based on variational mode decomposition, autoencoder and transfer learning to achieve high-precision regression prediction. Zhou et al. [16] proposed a joint distributed adaptive regression soft sensor model based on online fuzzy sets, which converted continuous labels into fuzzy class labels through fuzzy sets. By adapting both marginal and conditional distributions, the domain adaption of damage quantification task was realized, which significantly improved the accuracy of damage quantification in real-world environment. Zhu et al. [17] proposed an offset compensation Gaussian process regression model for the quality inference of chemical processes with distributed outputs. The molecular weight distribution prediction in a polymerization process indicates its feasibility and superiority. Liu et al. [18] proposed a novel framework of

an adversarial transfer learning (ATL) based soft sensing method for the quality inferring of multigrade processes. Liu et al. [19] proposed a soft sensing method based on domain adaptation extreme learning machine (DAELM), and the prediction results of two multilevel chemical processes showed the superiority of DAELM method. Therefore, transfer learning can effectively solve the problem of model failure caused by the application of traditional soft sensor models to multiple operating conditions, and transfer learning can transfer knowledge from multiple known fermentation conditions to help accomplish the target condition learning task, which effectively alleviates the problem of insufficient samples in small-sample fermentation process. The soft sensor model using the idea of transfer learning can solve the problem of multiple working conditions in modeling to a certain extent, but the soft sensor modeling process of multiple working conditions using traditional transfer learning is to align the distribution of the entire modeling datasets through transfer learning, without considering the local structure presented by the fermentation process due to the characteristics of nonlinear, multi-stage and other characteristics. During the transfer process, local information is easily ignored, resulting in the loss of some data information during feature mapping and the inability to maintain the original data structure. The established soft sensor model suffers from underfitting, and there is still room for improvement in the accuracy of the soft sensor model.

Based on this, this paper proposes a soft sensor modeling method based on the transfer modeling framework of substructure domain for multiple working conditions of the fermentation process of *Pichia pastoris*. Firstly, the source and target domain data are divided into substructure domains by the Gaussian mixture model (GMM) clustering algorithm, while the sub-source domains are weighted according to the distance between the sub-source and sub-target domains. Secondly, the optimal subspace domain adaptation method combining multiple metric strategies (maximum variance, manifold regularization and distribution discrepancy minimization) is used to obtain the optimal subspace projection of the sub-source domain and the sub-target domain, and the optimal projection is used to project the data of the sub-source domain and the sub-target domain into the manifold space to align the data of the two domains, so as to reduce the discrepancy between the data of different working conditions. Finally, considering the nonlinear and small-sample characteristics of *Pichia pastoris* fermentation process, the least squares support vector machine(LSSVM) is used as the basic modeling method, and the data after substructure domain transfer is used to train the prediction model. Taking the process of *Pichia pastoris* fermentation to generate inulinase as an example for validation, the simulation results show that the OSDA-LSSVM soft sensor model predicted the root-mean-square errors of *Pichia pastoris* concentration and inulinase concentration with a reduction of 48.7% and 54.9%, respectively, compared with the traditional LSSVM model, and it can effectively improve the accuracy of the soft sensor model under multiple working conditions.

Methods

The soft sensor model established based on the idea of transfer learning effectively solves the problem of the performance degradation of the soft sensor model caused by the mismatch of data distribution under multiple working conditions. However, when the process data presents local structure due to the nonlinearity and multi-stage characteristics, the transfer learning of the data as a whole means that the local information is ignored, and part of the data information will be lost during the feature mapping, so that the accuracy of the soft sensor model established on this basis will be affected as well. Therefore, it is necessary to improve the traditional transfer learning method and construct a framework for aligning data distribution within the local data structure with the highest relevance and then building the soft sensor model. In summary, taking into account the characteristics of biological fermentation process data, such as multiple working condition, multiple stage, and locality, this paper proposes a soft sensor modeling method based on the transfer modeling framework of substructure domain for multiple working conditions of the fermentation process of *Pichia pastoris*, as shown in Fig. 1.

Aiming at the multi-stage characteristics of the exogenous protein production process by *Pichia pastoris* fermentation, a transfer learning framework of substructure domain is introduced, in which the sample data are clustered to obtain the datasets (substructure domains) of each fermentation stage, and the substructures are weighted according to the distances between the substructures of the source and target domains, with larger weights indicating smaller discrepancies between the corresponding substructures of the source and target domains, and the weighted sub-source domains and the corresponding sub-target domains are taken as the novel sample data for the data transfer and transformation, which avoids the problem that the local prediction values of a single global soft sensor model deviate greatly from the actual values when estimating the key biological variables of the fermentation process, leading to an increase in the prediction error of the model. Simultaneously, considering the characteristics of multiple working conditions of the process of exogenous protein production

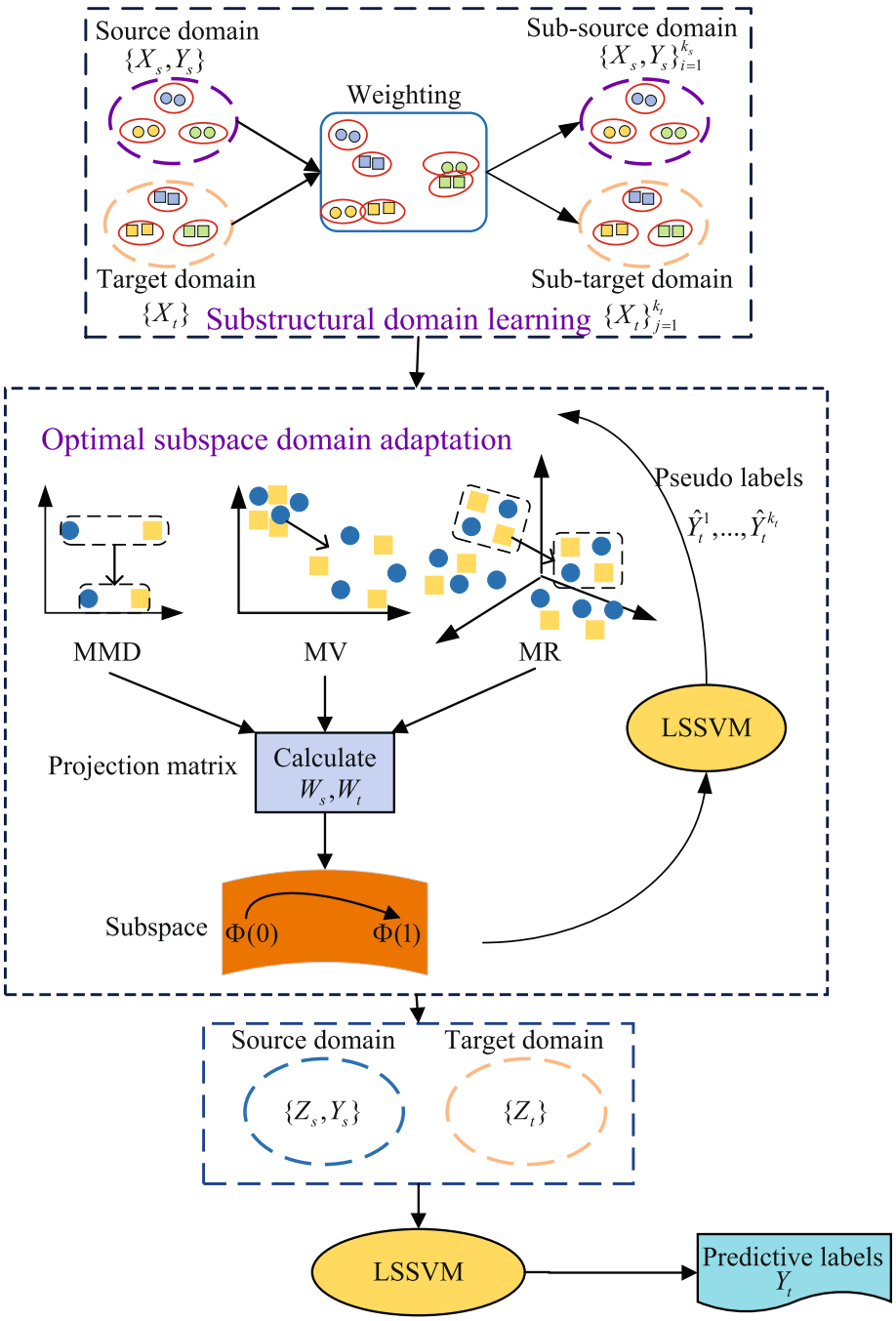

Fig. 1 The transfer learning framework of substructure domain based on OSDA-LSSVM

by *Pichia pastoris* fermentation, on the basis of substructural transfer, combining data transfer and subspace alignment, using multiple metric strategies (maximum variance, manifold regularization and distribution discrepancy minimization) to obtain the optimal projection matrices of the subspaces of the source and target domains, and projecting the data of each sub-source and sub-target domains into the subspaces for subspace alignment, which reduces the data distribution discrepancy among different working conditions and preserves the internal attributes and the neighborhood structure of the original data, to effectively solve the problem of model failure caused by multiple working conditions. Finally, the LSSVM prediction model is established using

the source and target domain data after substructure domain adaptation to realize the real-time prediction of the key biological variables in the production process of *Pichia pastoris* fermentation.

Substructure domain learning strategies

Aiming at the problem that the traditional transfer method is prone to lose data information in the overall transfer, and is difficult to achieve the perfect match between the two domains, thus reducing the accuracy of the soft sensor model, this paper introduces the transfer learning strategy of substructure domain transfer to achieve a more detailed substructure-level match between the two domains [20]. The substructure domain transfer strategy firstly clusters the source and target domain data using the GMM clustering algorithm to obtain the substructures (sub-source domains and sub-target domains) of the two domains. Then, the sub-source domains are adaptively weighted according to the distance between the sub-source domain and sub-target domain, and the weights represent the degree of similarity of the substructures in the two domains. Finally, mapping is performed between the substructures of the two domains, i.e., knowledge transfer is performed at each stage corresponding to the known working conditions and the working conditions to be measured, which in turn reduces the discrepancy between the two domains. Substructure-level transfer pays more attention to the transfer between substructures with small discrepancies, performs more fine-grained transfer learning between the sub-target domain and its most relevant sub-source domain, and avoids the noise introduced by domain-level transfer to a certain extent, so as to make better use of local information and improve the prediction accuracy of soft sensor.

Acquisition and representation of substructures

Using χ and $\delta \sim \mathcal{N}(0, \sigma^2)$ to represent the sample characterization data, χ^k conforms to $N(\varepsilon^k, \sigma^k)$, a Gaussian distribution. ε^k denotes the kth substructure center value, σ^k denotes the kth substructure covariance, and χ^k denotes that the data belongs to the kth substructure. The ε^k and σ^k can be obtained through χ. Considering the source and target domains as a mixture of GMM distributions, the Bayesian information criterion (BIC) is utilized to determine the number of substructures, i.e.

$$BIC = -2\ln(L) + h\ln(n) \tag{1}$$

Where, L represents the maximum value of the likelihood function of the estimated model, h represents the number of free parameters to be estimated, and n represents the sample size. The goal is to seek to minimize h of *BIC*.

After obtaining the substructure of the source domain and the target domain, the two domains can be represented as: $\tau_s = \sum_{i=1}^{k_s} w_i^s \delta_{\varepsilon_i^s}$, $\tau_t = \sum_{j=1}^{k_t} w_j^t \delta_{\varepsilon_j^t}$. This representation uses only the information of the cluster center, and the calculation is simple and efficient. Where ε represents the cluster center, δ_ε is the Dirac function at position ε, τ_s and τ_t are the distribution of the source and target domains respectively, and w is the probability associated with ε. Obviously, $\sum_{i=1}^{k_s} w_i^s = 1, \sum_{j=1}^{k_t} w_j^t = 1$. Here the square Euclidean distance is chosen as the cost between the source domain substructure ε_i^s and the target domain substructure ε_j^t, i.e.

$$C(\varepsilon_i^s, \varepsilon_j^t) = \left\| \varepsilon_i^s - \varepsilon_j^t \right\|^2 \tag{2}$$

Adaptive weighting of substructure based on optimal transmission

Since the target domain has less labeling information, the same weight is given to the substructure of the target domain, i.e., fixing w_j^t to $1/k_t$. It is known that $\sum_{i=1}^{k_s} w_i^s = 1$, the source domain substructure can be weighted by locally optimal transport with the following optimization objective.

$$\begin{aligned} &\pi_1^* = \arg\min_{\pi} \langle \pi, \mathbf{C} \rangle_F + \lambda H(\pi) \\ &s.t.\ \pi^T 1_{k_s} = w^t \end{aligned} \tag{3}$$

Where, $\langle \pi, \mathbf{C} \rangle_F$ is the total cost of the locally optimal transportation problem, $H(\pi) = \sum_{ij} \pi_{ij} \log \pi_{ij}$ is the entropy term, $\langle \cdot, \cdot \rangle_F$is the Frobenius dot product, $\mathbf{C}$ is the cost matrix, π is the coupling matrix between the two probability distribution functions, and λ is the hyperparameter of the balance calculation speed and precision.

Through Lagrange method, it is easy to obtain the optimal π_1^* as

$$\pi_1^* = \pi_0 diag(w^t \oslash \pi_0^T 1_{k_s}) \tag{4}$$

Where, $\pi_0 = e^{(-\mathbf{C}/\lambda)-1}$ is the result of the initialization, and $\oslash$ denotes division by elements. After obtaining the optimal coupling matrix π_1^*, the weight of each substructure of the source domain is $w^s = \pi_1^* 1_{k_t}$.

After obtaining the weighted source domain substructures and target domain substructures, mapping between the substructures, i.e., knowledge transfer at the substructure level, can be performed. Compared with the overall transfer learning at the domain level, the substructure-level transfer learning is more detailed and more in line with the multi-stage characteristics of the *Pichia pastoris* fermentation. The main process of substructural transfer learning is shown in Fig. 2.

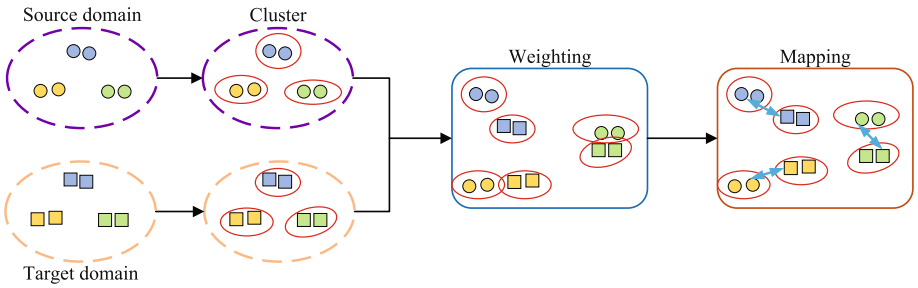

Fig. 2 Substructural transfer learning framework

Substructure mapping based on optimal subspace domain adaptation

Both traditional data centric and subspace centric domain adaptation methods have certain limitations. The data centric transfer learning method seeks a transformation matrix that minimizes the distance between the source and target domains in the common space, and due to the distribution discrepancy between the source and target domain data, there may not be such a common projection matrix. However, the subspace centric transfer learning method assumes that the source and target domain data have similar distribution in the transformed subspace, and the subspace alignment may fail when the discrepancy between the two domains is large.

Based on the above analysis, and considering the characteristics of industrial process data such as multiple working condition, multiple stage and locality, this paper proposes an optimal subspace domain adaptation (OSDA) method using the shared and domain-specific features of two domains. This method minimizes the distribution discrepancy between the two domains based on the improved balanced distribution adaptation algorithm, and introduces the maximum variance and manifold regularization methods to ensure that the projected data can retain the internal attributes and neighborhood structure of the original data, and seeks the two mutually coupled optimal projections. Secondly, the optimal projection matrices are used to replace the traditional projection matrices (obtained by principal component analysis (PCA) used by the geodetic flow kernel (GFK) method) to project the source and target domain data into the source and target domain subspaces, respectively, and further align the subspaces so as to reduce the discrepancy between the source and target domain data. The OSDA method combines the data centric and subspace centric methods, and reduces the discrepancies of different batches of *Pichia pastoris* fermentation data in terms of statistics and geometry structure, which makes the established soft sensor model applicable to new working conditions and improves the generalization ability of the soft sensor model.

Assumed a labeled source domain sample $D_s = \{x_{si}, y_{si}\}_{i=1}^{n_s}$ and a less labeled or unlabeled target domain sample $D_t = \{x_{tj}\}_{j=1}^{n_t}$ of *Pichia pastoris* fermentation. The source domain feature data is denoted as $X_s \in \mathbb{R}^{d \times n_s}$, and the target domain feature data is denoted as $X_t \in \mathbb{R}^{d \times n_t}$, where d represents the sample feature dimension, and n_s and n_t represent the number of samples in the source domain and target domain respectively. Assume that the feature space and label space of the two domains are the same, i.e., $\mathcal{X}_s = \mathcal{X}_t$, $\mathcal{Y}_s = \mathcal{Y}_t$. But the marginal probability distribution and conditional probability distribution are different, i.e., $P(x_s) \neq P(x_t)$ and $P(y_s|x_s) \neq P(y_t|x_t)$.

Optimal subspace acquisition

In traditional transfer learning, balanced distribution adaptation (BDA) method is mainly used to solve the problem of process data distribution matching. BDA adapts the marginal distribution and conditional distribution between two domains via maximum mean discrepancy (MMD), thereby reducing the discrepancy in probability distribution between the two domains [21, 22]. Marginal distribution adaptation calculates the distance between the sample mean of the source domain and the target domain in the low-dimensional embedding, so that the marginal probability distributions of the two domains are approximately equal after projection, i.e., $P({W_s}^T x_s) \approx P({W_t}^T x_t)$. Conditional distributed adaptation utilizes the class conditional probability to approximate the conditional probability, trains a classifier through source domain data to obtain the target domain pseudo-label $\hat{Y}_t$, and iterates times T to improve the accuracy of the pseudo-label. Conditional distribution adaptation calculates the distance between the sample means of all classes, such that $P(y_s|{W_s}^T x_s) \approx P(y_t|{W_t}^T x_t)$. The discrepancy in

probability distributions between the two domains is defined as follows.

$$D(X_s, X_t) \approx (1-\eta)\left\| \frac{1}{n_s}\sum_{x_i \in X_s} {W_s}^T x_i - \frac{1}{n_t}\sum_{x_j \in X_t} {W_t}^T x_j \right\|_F^2 + \eta \sum_{c=1}^{C} \left\| \frac{1}{n_s^{(c)}}\sum_{x_i \in X_s^{(c)}} {W_s}^T x_i - \frac{1}{n_t^{(c)}}\sum_{x_j \in X_t^{(c)}} {W_t}^T x_j \right\|_F^2 \quad (5)$$

Where, η is a balance factor and $\eta \in [0,1]$, is used to dynamically adjust the importance of the marginal and conditional distributions. $n_s^{(c)}$ and $n_t^{(c)}$ denote the number of samples belonging to class c in the source and target domains, and $X_s^{(c)}$ and $X_t^{(c)}$ denote the samples belonging to class c in the source and target domains, and W_s and W_t are projection matrices that project the source domain and the target domain into the subspace, respectively.

MMD conditional distribution adaptation is to use class conditional probability to approximate conditional probability, while the soft sensor modeling process of biochemical reaction process belongs to the regression task, and its labels are continuous. If BDA method is used for transfer learning, continuous labels need to be constrained into classes to obtain "class labels", and then conditional distribution adaptation is realized.

Based on the above, this paper introduces the concept of fuzzy set [23], and restricts the continuous labels in the fermentation process to the fuzzy class through fuzzy set, i.e., the values at 5%, 50% and 95% of the continuous labels in the source and target domains are taken as the class center of the fuzzy class. As shown in Fig. 3a, each continuous source domain label can belong to three fuzzy classes of small^s, medium^s and large^s at the same time.

The class of small^s is taken as the first class, the class of medium^s as the second class, and the class of large^s as the third class, and the membership degree μ_{ic}^s indicates the extent to which the source domain label y_i^s belongs to the class c. The membership degree is normalized according to μ_{ic}^s for each class. i.e.

$$\bar{\mu}_{ic}^s = \frac{\mu_{ic}^s}{\sum_{i=1}^{n_s} \mu_{ic}^s}, i = 1, ..., n_s; c = 1, 2, 3 \quad (6)$$

Similarly, three fuzzy classes of the target domain pseudo-label can be obtained, as shown in Fig. 3b. Its membership degree is:

$$\bar{\mu}_{jc}^t = \frac{\mu_{jc}^t}{\sum_{j=1}^{n_t} \mu_{jc}^t}, j = 1, ..., n_t; c = 1, 2, 3 \quad (7)$$

According to Eqs. 5, 6 and 7, the updated distribution discrepancy is defined as:

$$D(X_s, X_t) \approx (1-\eta)\left\| \frac{1}{n_s}\sum_{x_i \in X_s} {W_s}^T x_i - \frac{1}{n_t}\sum_{x_j \in X_t} {W_t}^T x_j \right\|_F^2 + \eta \sum_{c=1}^{3} \left\| \sum_{x_i \in X_s} \bar{\mu}_{ic}^s {W_s}^T x_i - \sum_{x_j \in X_t} \bar{\mu}_{jc}^t {W_t}^T x_j \right\|_F^2 \quad (8)$$

Introducing the kernel trick, the distribution discrepancy function is rewritten as follows.

$$\min_W Tr\left(W^T S_{mmd} W\right) \quad (9)$$

$$\text{Where, } W = \begin{bmatrix} W_s \\ W_t \end{bmatrix}, \quad S_{mmd} = \begin{bmatrix} M_s & M_{st} \\ M_{ts} & M_t \end{bmatrix} \quad (10)$$

$$M_s = X_s\left((1-\eta)N_s + \eta\sum_{c=1}^{3} N_s^{(c)}\right){X_s}^T, N_s = \frac{1}{n_s^2} 1_s 1_s^T, \quad (N_s^{(c)})_{ij} = \begin{cases} \bar{\mu}_{ic}^s \bar{\mu}_{jc}^s & x_i, x_j \in X_s^{(c)} \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

$$M_t = X_t\left((1-\eta)N_t + \eta\sum_{c=1}^{3} N_t^{(c)}\right){X_t}^T, N_t = \frac{1}{n_t^2} 1_t 1_t^T, \quad (N_t^{(c)})_{ij} = \begin{cases} \bar{\mu}_{ic}^t \bar{\mu}_{jc}^t & x_i, x_j \in X_t^{(c)} \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

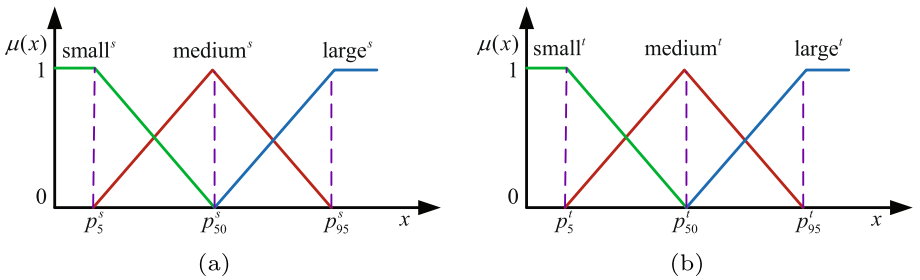

Fig. 3 Fuzzy class division

$$M_{st} = X_s\left((1-\eta)N_{st} + \eta\sum_{c=1}^{3}N_{st}^{(c)}\right){X_t}^T, N_{st} = -\frac{1}{n_s n_t}1_s 1_t^T,$$
$$(N_{st}^{(c)})_{ij} = \begin{cases} -\bar{\mu}_{ic}^s\bar{\mu}_{jc}^t & x_i \in X_s^{(c)}, x_j \in X_t^{(c)} \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

$$M_{ts} = X_t\left((1-\eta)N_{ts} + \eta\sum_{c=1}^{3}N_{ts}^{(c)}\right){X_s}^T, N_{ts} = -\frac{1}{n_s n_t}1_t 1_s^T,$$
$$(N_{ts}^{(c)})_{ij} = \begin{cases} -\bar{\mu}_{ic}^t\bar{\mu}_{jc}^s & x_i \in X_t^{(c)}, x_j \in X_s^{(c)} \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

Meanwhile, to ensure the ability to represent different features of the source domain and the target domain, and avoid projecting the features of the source domain and the target domain into unrelated dimensions, this paper introduces the maximum variance (MV) [24]. The optimization objective is set as:

$$\max_W Tr\left(W^T S_{mv} W\right) \tag{15}$$

Where,

$$S_{mv} = \begin{bmatrix} V_s & 0 \\ 0 & V_t \end{bmatrix} \tag{16}$$

$$V_s = X_s H_s X_s^T \tag{17}$$

$$V_t = X_t H_t X_t^T \tag{18}$$

Where, $H_s = I_s - \frac{1}{n_s}1_s 1_s^T$ and $H_t = I_t - \frac{1}{n_t}1_t 1_t^T$ are both central matrices, $I_s \in \mathbb{R}^{n_s \times n_s}$ and $I_t \in \mathbb{R}^{n_t \times n_t}$ are identity matrices, and $1_s \in \mathbb{R}^{n_s}$ and $1_t \in \mathbb{R}^{n_t}$ are all-one column vectors.

Moreover, in order to further maintain the structural information of the source and target domains during the projection process, manifold regularization (MR) is introduced to extract the local neighborhood features of the data through MR, and maintain this structure in the manifold space after the projection [25, 26]. Its objective function is:

$$R_f(X_s, X_t) = \sum_{i,j=1}^{n_s+\mathrm{n}_t} G_{ij}\left\| W_s^T x_i - W_t^T x_j \right\|_F^2 \tag{19}$$

Where, $G_{ij} = e^{-\|x_i - x_j\|^2/t}$ denotes the similarity between the two sample points x_i and x_j, and the final regularization can be written as:

$$\min_W Tr\left(W^T S_{mr} W\right) \tag{20}$$

Where,

$$S_{mr} = \begin{bmatrix} R_s & R_{st} \\ R_{ts} & R_t \end{bmatrix} \tag{21}$$

$$R_s = X_s L_s X_s^T \tag{22}$$

$$R_{st} = X_s L_{st} X_t^T \tag{23}$$

$$R_{ts} = X_t L_{ts} X_s^T \tag{24}$$

$$R_t = X_t L_t X_t^T \tag{25}$$

Where, $L = D - G$ is the Laplacian matrix and $D_{ii} = \sum_{j=1}^{n_s+n_t} G_{ij}$ is the diagonal matrix.

The improved OSDA greatly reduces the discrepancy between the source and target domain subspaces by simultaneously optimizing W_s and W_t to be close to the source and target domain subspaces.

To control the size of the projection matrix, regular constraints $\|W_s\|_F^2$ and $\|W_t\|_F^2$ are further introduced. The objective function is set as follows.

$$\min_{W_s, W_t} \| W_s - W_t \|_F^2 + \| W_s \|_F^2 + \| W_t \|_F^2 \tag{26}$$

Combining Eqs. 9, 15, 20 and 26, the objective function is obtained as follows.

$$\max \frac{\theta_1\{MV\}}{\{MMD\} + \theta_2\{MR\} + \theta_3 \| W_s - W_t \|_F^2 + \alpha \| W_s \|_F^2 + \beta \| W_t \|_F^2} \tag{27}$$

Where, θ_1 , θ_2 and θ_3 are balancing parameters that balance the importance of each quantity and take values in the range of [0, 1], and α and β are the regular coefficients. Combining Eqs. 10, 16 and 21 rewrites Eq. 27 as:

$$\max_W Tr\left(W^T \begin{bmatrix} \theta_1 V_s & \\ & \theta_1 V_t \end{bmatrix} W\right)$$
$$s.t.\ Tr\left(W^T \begin{bmatrix} M_s + \theta_2 R_s + (\theta_3 + \alpha)I & M_{st} + \theta_2 R_{st} - \theta_3 I \\ M_{ts} + \theta_2 R_{ts} - \theta_3 I & M_t + \theta_2 R_t + (\theta_3 + \beta)I \end{bmatrix} W\right) = 1 \tag{28}$$

By the Lagrangian method, it is finally obtained:

$$\begin{bmatrix} \theta_1 V_s & \\ & \theta_1 V_t \end{bmatrix} W = \begin{bmatrix} M_s + \theta_2 R_s + (\theta_3 + \alpha)I & M_{st} + \theta_2 R_{st} - \theta_3 I \\ M_{ts} + \theta_2 R_{ts} - \theta_3 I & M_t + \theta_2 R_t + (\theta_3 + \beta)I \end{bmatrix} W\phi \tag{29}$$

Where, $\phi = diag(\lambda_1, ..., \lambda_m)$ is the first m eigenvalues, $W = (W_1, ..., W_m)$ contains the corresponding eigenvectors, which can be solved by generalized eigenvalue decomposition, and finally the optimal projection matrices W_s and W_t are obtained.

Subspace alignment

Consider the optimal projection matrices W_s and W_t as two points in the manifold space, such that $W_s = \Phi(0)$, $W_t = \Phi(1)$, and a geodesic $\{\Phi(t) : 0 \le t \le 1\}$ between the two points can form a path between the two subspaces. The phenomenon of drift between domains is reduced by finding a geodesic line from $\Phi(0)$ to $\Phi(1)$.

The features in the transformed manifold space can be denoted as $z = \Phi(t)^T x$. The transformation from $\Phi(0)$ to $\Phi(1)$ passes through several points, which is accomplished by defining a semi-positive definite geodetic flow kernel through the inner product of the transformed features.

$$\langle z_i, z_j \rangle = \int_0^1 \left(\Phi(t)^T x_i\right)^T \left(\Phi(t)^T x_j\right) dt = x_i^T G x_j \tag{30}$$

The source and target domain data after subspace alignment are: $Z_s = \sqrt{G} X_s$, $Z_t = \sqrt{G} X_t$.

Different from the traditional GFK method, OSDA uses multiple metric strategies (maximum variance, manifold regularization and distribution discrepancy minimization) to obtain the optimal projection matrices of the source and target domain subspaces, and uses the optimal projection matrices to replace the S_s and S_t projection matrices obtained by PCA method in GFK, which better realizes the alignment of the source domain subspaces and target domain subspaces.

Least squares support vector machine

Considering that the least squares support vector machine (LSSVM) has better performance in solving small sample and nonlinear problems, this paper adopts the source and target domain data after substructure domain adaptation to train the LSSVM, and constructs the soft sensor model of the production process of *Pichia pastoris* fermentation.

LSSVM is a novel type of support vector machine proposed by Suykens on the basis of support vector machine for solving model decomposition and function estimation problems [27]. Suppose there are l training samples $\{(x_i, y_i)|i = 1, 2, ..., l\}$, in which the samples are n-dimensional vectors, $x_i \in \mathbb{R}^n$ is the sample input, $y_i \in \mathbb{R}^n$ is the sample output, and the optimization objective of LSSVM is:

$$\begin{aligned} &\min_{\omega,b,\xi} J(\omega, \xi) = \frac{1}{2}\omega^T \omega + \frac{1}{2}\gamma \sum_{i=1}^{l} \xi_i^2 \\ &s.t.\ y_i = \omega^T \varphi(x_i) + b + \xi_i (i = 1, 2, \cdots, l) \end{aligned} \tag{31}$$

Where, ω is the weight vector, ξ_i is the error variable, b is the deviation quantity, γ is the penalty coefficient, and $\varphi(\cdot)$ is the nonlinear mapping.

The final function is estimated by the Lagrangian method to solve for:

$$f(x) = \sum_{i=1}^{l} \alpha_i K(x, x_i) + b \tag{32}$$

Where, $K(x, x_i)$ is the kernel function, which has various forms such as radial basis function (RBF) and polynomial function. In this paper, RBF is used as the kernel function. For the two hyperparameters that affect the performance of LSSVM model, the regular coefficient and kernel width, this paper simply combines the fast leave-one-out cross-validation method to optimize the regular coefficient and RBF kernel width.

Soft sensor modeling based on OSDA-LSSVM

Considering the characteristics of *Pichia pastoris* fermentation, such as multiple working condition, multiple stage and locality, this paper transfers and transforms the fermentation process data based on the transfer modeling framework of substructure domain, and constructs a soft sensor model of the fermentation process based on the LSSVM modeling method with simple structure and strong generalization ability. In addition, we verify the performance of the soft sensor model in the simulation environment of MATLAB 2017a (with LSSVM support package added).

The specific steps of soft sensor modeling method based on the transfer modeling framework of substructure domain are as follows:

Step1: The sample data of *Pichia pastoris* fermentation experiment were obtained, and the sample datasets were established, and the datasets were preprocessed. According to the consistent correlation degree method, auxiliary variables with correlation degree greater than 0.8 were selected to construct the data of source domain $D_s = \{X_s, Y_s\}$ and target domain $D_t = \{X_t\}$.

Step2: The substructures of source and target domain data are obtained by GMM clustering algorithm, and the sub-source domains are adaptively weighted according to the distance between the sub-source and sub-target domains.

Step3: The weighted sub-source domains $D_s^1, D_s^2, ..., D_s^{k_s}$ and the corresponding sub-target domains $D_t^1, D_t^2, ..., D_t^{k_t}$ are used as novel sample data to train

the LSSVM model, and the pseudo labels $\hat{Y}_t^1, \hat{Y}_t^2, ..., \hat{Y}_t^{k_t}$ of the sub-target domains are obtained.

Step4: The sub-source domains and sub-target domains with pseudo labels are taken as new sample data, and the optimal projection matrices W_s and W_t are calculated by the optimal subspace domain adaptation method integrating multiple metric strategies (maximum variance, manifold regularization and distribution discrepancy minimization). Then, the data of each sub-source domain and sub-target domain are projected into the subspace to further realize the transformation from sub-source domain to sub-target domain, and obtain the data of source domain and target domain after reducing the distribution discrepancy.

Step5: The LSSVM soft sensor model is built using the source domain data $\{Z_s, Y_s\}$ and target domain data $\{Z_t\}$ after substructure domain adaptation to obtain the actual predicted labelsY_t.

In summary, Algorithm 1 shows more specific steps for OSDA-LSSVM.

Algorithm 1 Optimal Subspace Domain Adaptation based on LSSVM

Input: Feature data and source domain labels: X_s, X_t, Y_s; Parameters: λ, η, $\theta_1 = 1$, $\theta_2 = 1$, $\theta_3 = 1$, α, β, T, m.

Output: Transformation matrices: W_s and W_t; Soft sensor model: f; Actual predictive labels: Y_t.

1: The substructures of source and target domains are obtained by GMM clustering;
2: The cost matrix $\mathbf{C}$ is calculated according to Eq. 2, and the weights w^s of the source domain substructures are calculated using Eqs. 3 and 4 to obtain the weighted substructure representations $\{\tilde{X}_s, \tilde{Y}_s\}_{i=1}^{k_s}$ and $\{\tilde{X}_t\}_{j=1}^{k_t}$;
3: Calculate S_{mmd}, S_{mv} and S_{mr} according to Eqs. 10, 16 and 21, and initialize the pseudo labels $\{\hat{Y}_t\}_{j=1}^{k_t}$ in the sub-target domain using the LSSVM trained on the sub-source domain data;
4: **repeat**
5: The generalized eigenvalue decomposition is used to solve Eq. 29, and the m eigenvectors corresponding to the first m eigenvalues are chosen as the transformation $\tilde{W}$ to get the projections $\tilde{W}_s$ and $\tilde{W}_t$;
6: Train the LSSVM soft sensor model $\tilde{f}$ based on $\{\tilde{W}_s^T\tilde{X}_s, \tilde{Y}_s\}_{i=1}^{k_s}$ to update the pseudo labels $\hat{Y}_t = \tilde{f}(\tilde{W}_t^T\tilde{X}_t)$ of sub-target domain;
7: Update S_{mmd} according to Eq. 10;
8: **until** Convergence;
9: GFK is utilized to further align $\tilde{W}_s$ and $\tilde{W}_t$ to obtain the projections W_s and W_t of the entire source and target domains, as well as the novel sample features $\{Z_s\}$ and $\{Z_t\}$;
10: Train the soft sensor model f at $\{Z_s, Y_s\}$ and use $\{Z_t\}$ to get the actual prediction labels Y_t.

Important biochemical variables involved in the fermentation process of *Pichia pastoris* include *Pichia pastoris* concentration, methanol concentration and inulinase concentration, among which the methanol concentration can be measured online by the corresponding laboratory-level analytical instrument or meters, while the *Pichia pastoris* concentration and inulinase concentration can only be obtained by offline and laboratory analysis in more cases, which not only costs a lot of manpower and material resources, but also affects the implementation of fermentation process control strategy and the improvement of fermentation technology. Based on this, this paper constructs the soft sensor model of the key biological variables (*Pichia pastoris* concentration and inulinase concentration) in the process of inulinase production by *Pichia pastoris* fermentation based on the transfer modeling framework of substructure domain, to provide important information for the online control and optimization of the process of inulinase production by *Pichia pastoris* fermentation.

Pichia pastoris GS115, MutsHis+ strain was selected for methanol-induced expression of inulin endonuclease INU2 on the transformants, and the enzyme activity of recombinant inulinase was detected. The inulinase generation process test platform was provided by Yangzhong Fructooligosaccharide(FOS) has been widely used in the field of health food because of its indigestibility, low caries coelicity and improving lipid metabolism. At present, one of the ways to prepare FOS is to hydrolyze inulin with endo-inulinase produced by *Pichia pastoris* fermentation. Weikert Bioengineering Equipment Co., Ltd, and the RTY0-C-100L fermenter was used as the fermentation equipment. The process of inulinase generation by *Pichia pastoris* fermentation is shown in Fig. 4.

In order to make the experiment close to the actual production process, the experimental process is designed as follows:

1. According to the requirements of *Pichia pastoris* strain inoculation, preparation of medium, shaking bottle culture and sterilization of fermentation equipment were carried out. The medium was then sterilized at 130°C for 30 minutes. When the temperature dropped to 30°C, the strain was introduced into the fermenter by flame inoculation method. The initial fermentation conditions are shown in Table 1.
2. The auxiliary variables sampled every 15 minutes were archived in a structured database. *Pichia pastoris* concentration and inulinase concentration were sampled offline every two hours and recorded. The data pairs of auxiliary variables and biological variables were established by interpolation method as the fermentation sample data of this batch. We selected fermentation broth temperature (T), pH, dissolved oxygen concentration (Do), stirring rate (r), and intake flow rate (V) as auxiliary variables.
3. The fermentation cycle of *Pichia pastoris* is 90 hours, and each batch contains 180 sample data. The auxiliary variables were taken as inputs, *Pichia pastoris* concentration and inulinase concentration as outputs, which were combined with the established soft sensor model to realize the real-time prediction of key biological variables.

Table 1 Initial fermentation conditions of *Pichia pastoris*

Environment variables	Initial setting value
Fermenter pressure	0.04*Mpa*
Motor stirring speed	300 400r/min
Ventilation capacity	150-300L/M
Fermentation liquid temperature	28±1°C
PH	5.0±0.4

To verify whether each strategy has a positive effect on the soft sensor model, we establish soft sensor models that remove a certain strategy and compare it with the model proposed in this paper. As shown in Fig. 5a, when the MV strategy is removed(Model1, i.e., $\theta_1 = 0$), the predicted value of the model for *Pichia pastoris* concentration begins to deviate greatly from the actual value, and as shown in Table 2, compared with

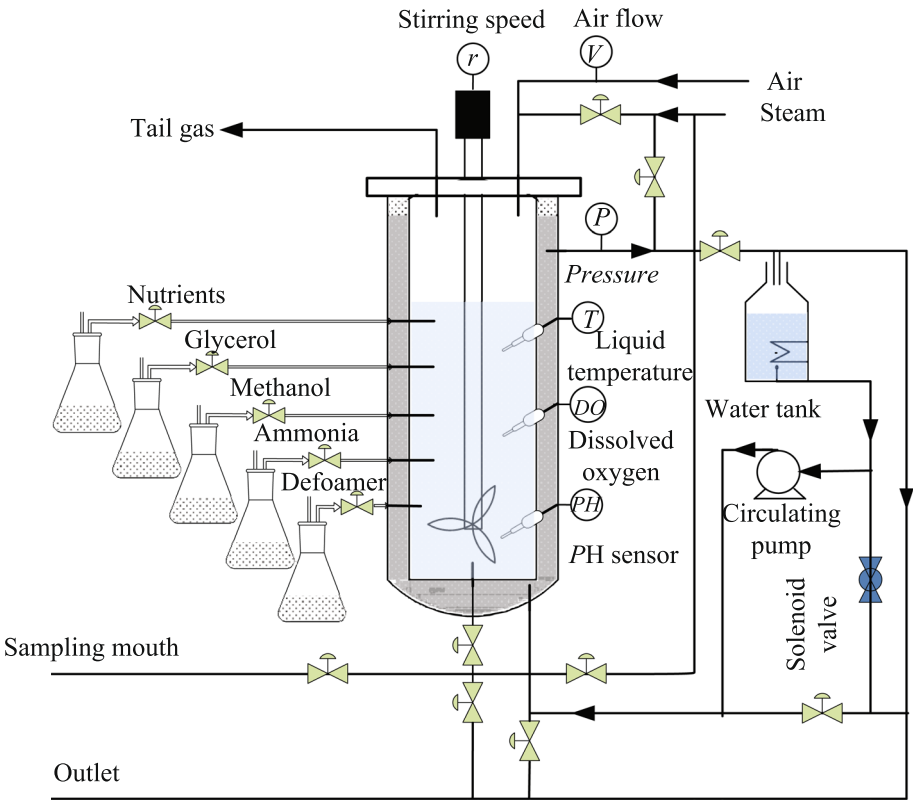

Fig. 4 Diagram of the process of *Pichia pastoris* fermentation to produce inulinase

the original OSDA-LSSVM model, the root mean square error increases, the coefficient of determination decreases, and the model performance decreases, which indicates that the MV strategy is very necessary. Similarly, Fig. 5b shows that the performance of the model is reduced to a certain extent when the MR Strategy is removed(Model2, i.e., $\theta_2 = 0$), and Table 2 also shows the important role of the MR Strategy in the soft sensor model. In addition, when the $\| W_s - W_t \|$ term is removed(Model3, i.e., $\theta_3 = 0$), as shown in Fig. 5c, the performance of the model decreases slightly, and it can be seen from Table 2 that this strategy has little effect on the soft sensor model. When the subspace alignment carried out by GFK method is removed(Model4), it can be seen from Fig. 5d and Table 2 that the performance of the model decreases greatly. In conclusion, MV, MR and GFK play an important role in soft sensor modeling methods.

Table 2 The assessment metrics of the comparison experiment of each strategy module

Soft sensor model	*Pichia pastoris* concentration	
	RMSE	R^2
OSDA-LSSVM	0.9560	0.9936
Model1	2.5049	0.9560
Model2	1.6229	0.9815
Model3	1.2229	0.9895
Model4	1.5302	0.9836

To verify the validity of the soft sensor modeling method proposed in this paper, the key biological variables(*Pichia pastoris* concentration and inulinase concentration) were predicted based on the constructed

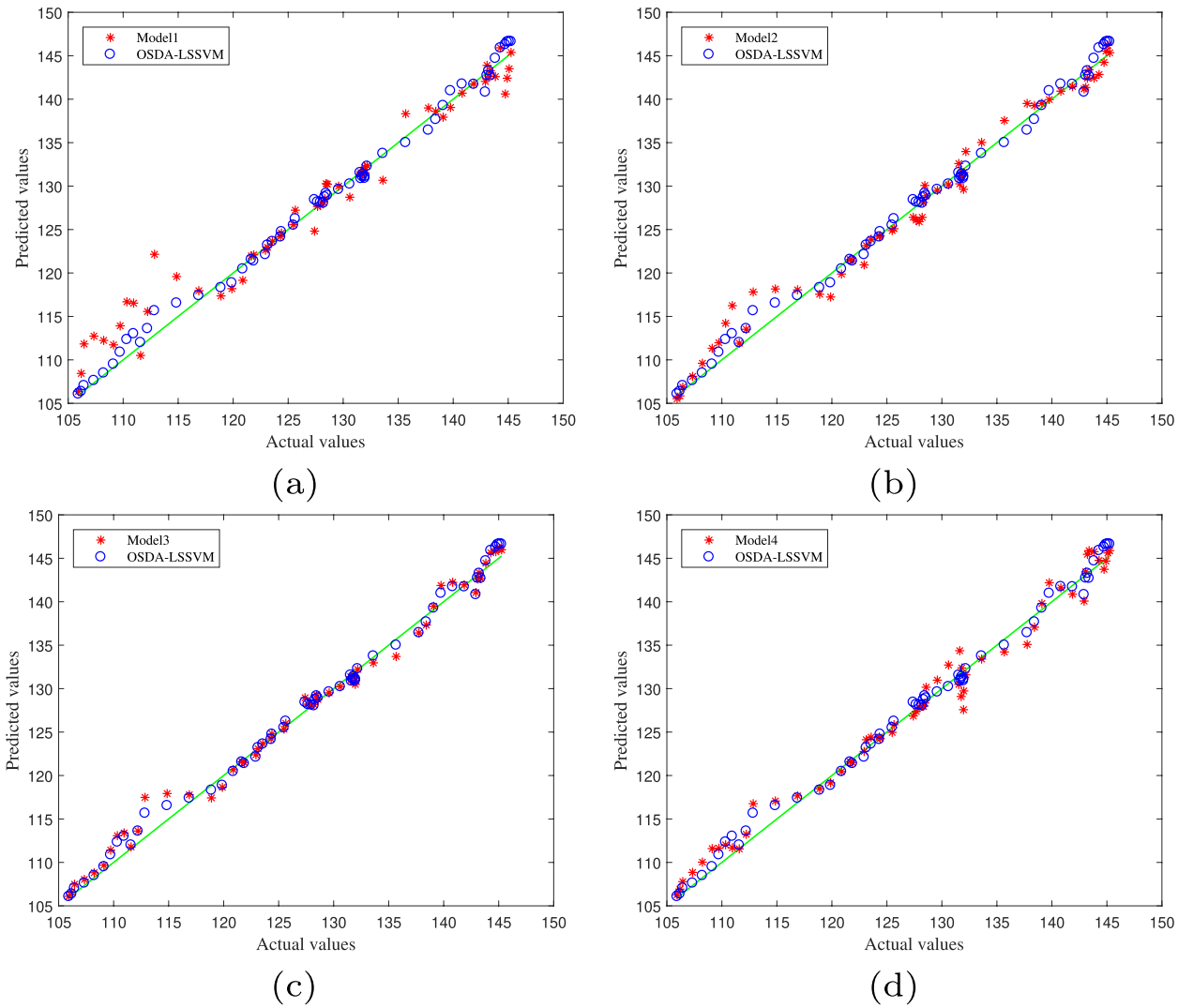

Fig. 5 Comparison experiment of the influence of each strategy module on the model

OSDA-LSSVM soft sensor model. Meanwhile, in order to verify the superior performance of the OSDA-LSSVM soft sensor model, this paper also established the LSSVM, GFK-LSSVM, BDA-LSSVM and OSDA-LSSVM soft sensor models based on the same batch of data. The prediction curves of the four soft sensor models for *Pichia pastoris* concentration are shown in Figs. 6 and 7 illustrates the curves of each of the four models to track the actual value, where the "Actual Value" is the *Pichia pastoris* concentration value sampled offline.

The LSSVM model in Fig. 7 uses RBF and adopts the reservation-one parameter algorithm to optimize the two hyperparameters of kernel function width and regularization coefficient. By comparing the prediction results of the LSSVM and GFK-LSSVM models in Fig. 7, it can be seen that there is a significant deviation in the overall prediction curve of the LSSVM model that uses the traditional reservation-one parameter algorithm to optimize hyperparameters. The GFK-LSSVM model introduces the subspace alignment algorithm in transfer learning, projects the sample data into the manifold space through the projection obtained by PCA, realizes the transformation of the training sample to the test sample, and thus improves the performance of the model. However, the local predicted value of GFK-LSSVM model deviates greatly from the actual value.

Compared with the GFK-LSSVM model, the BDA-LSSVM model combined with the BDA in transfer learning seeks a transformation that minimizes the discrepancy between the probability distribution of the training data and the test data in the common space, so that the prediction curve of the model is more consistent with the actual value. According to the results of many experiments, when the balance factor η of adjusting the marginal probability distribution and conditional probability distribution in BDA is set to 0.6, the BDA-LSSVM model has superior prediction performance.

Compared to the BDA-LSSVM model, the OSDA-LSSVM soft sensor model based on the transfer modeling framework of substructure domain proposed in this paper reduces the discrepancies of different batches of *Pichia pastoris* fermentation data in terms of statistics and geometric structure, so it can make full use of the local information of fermentation process data, and has higher prediction accuracy than the overall transfer soft sensor modeling, which can effectively improve the accuracy of soft sensor model under multiple working conditions. During the simulation process, we set the parameter $\eta = 0.6$ in OSDA to balance the two probability distributions, and set the balance parameters $\theta_1 = 1$, $\theta_2 = 1$ and $\theta_3 = 1$, i.e., the default is equally important. The number of iterations $T = 10$, the dimension of the final projection matrix $m = 5$. As can be seen from Fig. 7,

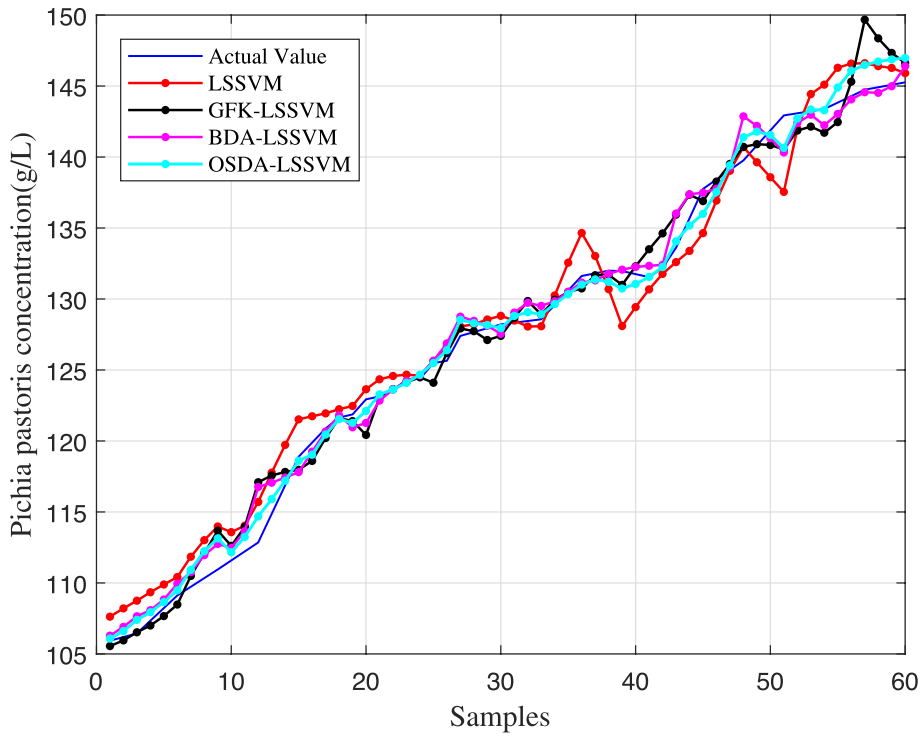

Fig. 6 Prediction curves for *Pichia pastoris* concentration

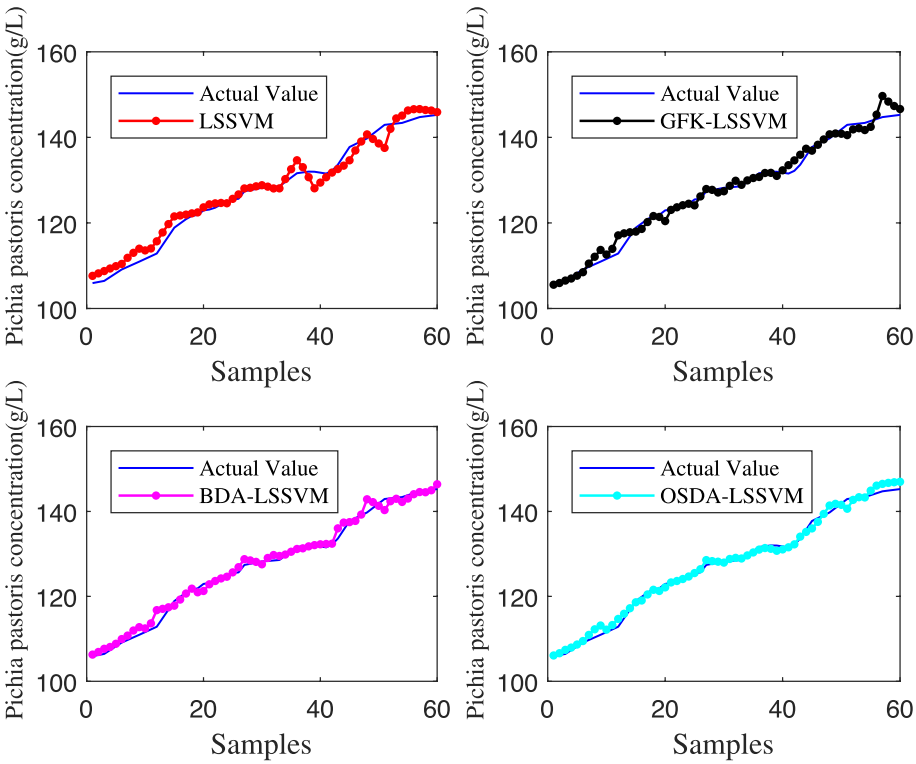

Fig. 7 Prediction curves of different models for *Pichia pastoris* concentration

the performance of OSDA-LSSVM soft sensor model is further improved, which can achieve real-time online accurate prediction of *Pichia pastoris* concentration.

In order to further verify the performance of the OSDA-LSSVM soft sensor model, the inulinase concentration in *Pichia pastoris* fermentation process is predicted based on the LSSVM, GFK-LSSVM, BDA-LSSVM and OSDA-LSSVM soft sensor models. As shown in Figs. 8 and 9, the simulation results show that the OSDA-LSSVM model also has superior performance in tracking and predicting inulinase concentration compared with the other three models, and its prediction curve can basically fit the actual value of inulinase concentration.

The relative error curves for *Pichia pastoris* concentration and inulinase concentration demonstrate more directly the predictive performance of the four soft sensor models, as shown in Figs. 10 and 11. Simulation results show that the proposed OSDA-LSSVM model has the smallest error.

To comprehensively compare the prediction effects of the four soft sensor models, this paper uses the root mean square error (RMSE) and coefficient of determination (R^2) to evaluate the prediction ability of the four soft sensor models, as shown in Table 3.

As can be seen from Table 3, compared with the other three models, the OSDA-LSSVM model has the smallest RMSE in predicting *Pichia pastoris* concentration and inulinase concentration, and its R^2 is closer to 1. To further verify the universality of the proposed model, the performance of the model is verified on another validation set, as shown in the Figs. 12 and 13. The results show that the OSDA-LSSVM soft sensor model has better generalization ability and higher prediction accuracy under multiple working conditions, and can better deal with the nonlinearity, time-varying and coupling of *Pichia pastoris* fermentation process.

Discussion

To address the limitations of single global model and traditional domain-level transfer learning method, this paper introduces the transfer learning strategy of substructure domain adaptation to achieve more detailed substructure-level matching between the two domains, extracts the local information of the *Pichia pastoris* fermentation process by Gaussian mixture model clustering algorithm, clusters the source and target domain data into multiple substructure domains, and constructs a local transfer framework to improve the model prediction performance. Meanwhile, on the basis of data transfer, combined with the method of subspace alignment, instead of seeking a common subspace with the smallest discrepancy, it seeks the respective subspaces of the two

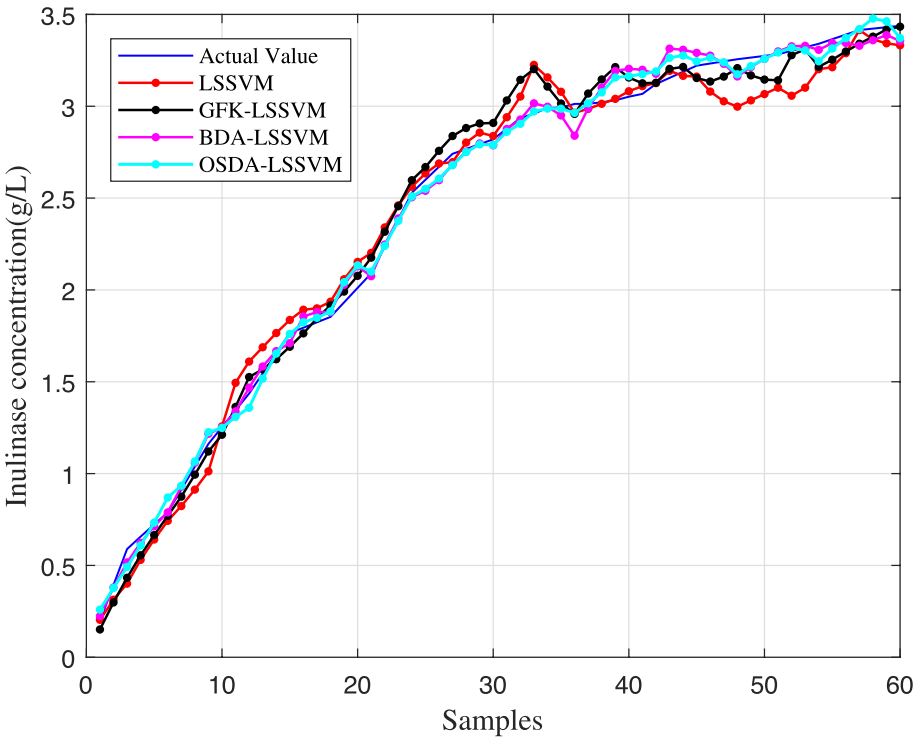

Fig. 8 Prediction curves for inulinase concentration

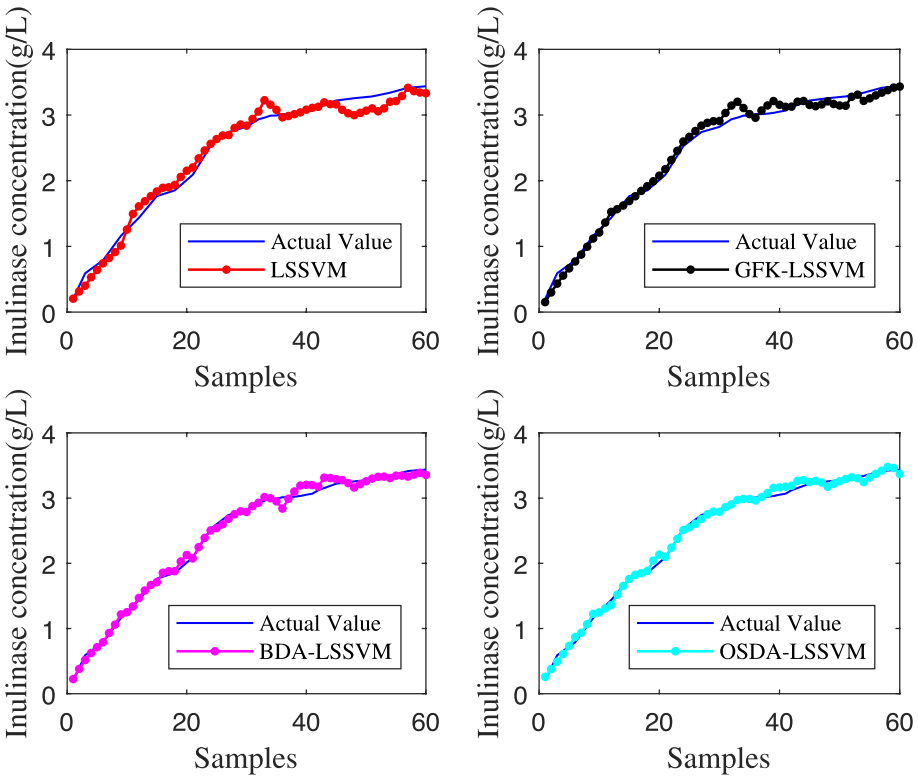

Fig. 9 Prediction curves of different models for inulinase concentration

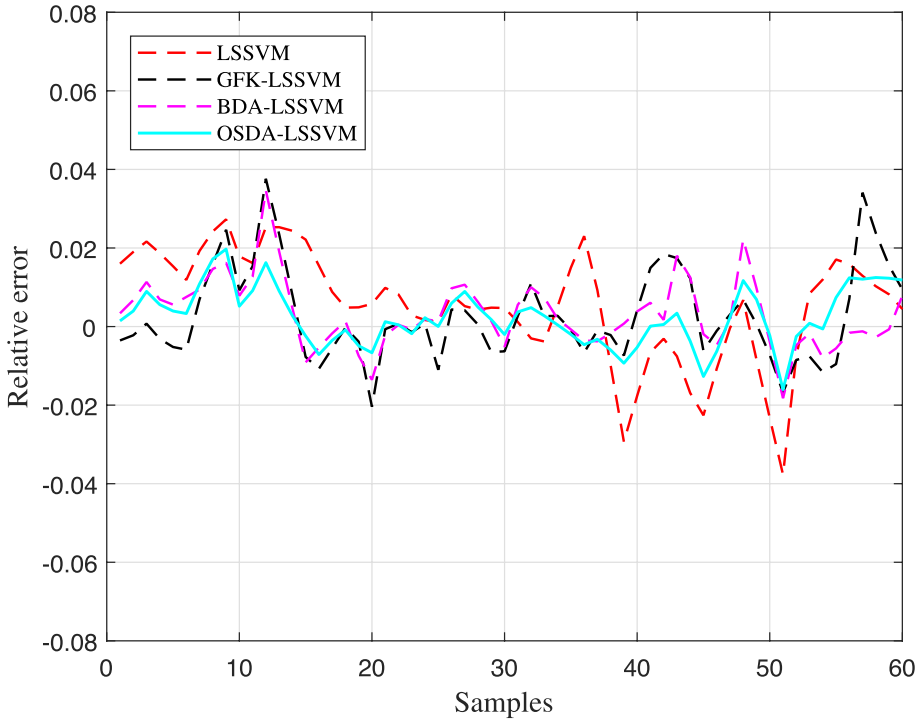

Fig. 10 Relative error curves of different soft sensor models in predicting *Pichia pastoris* concentration

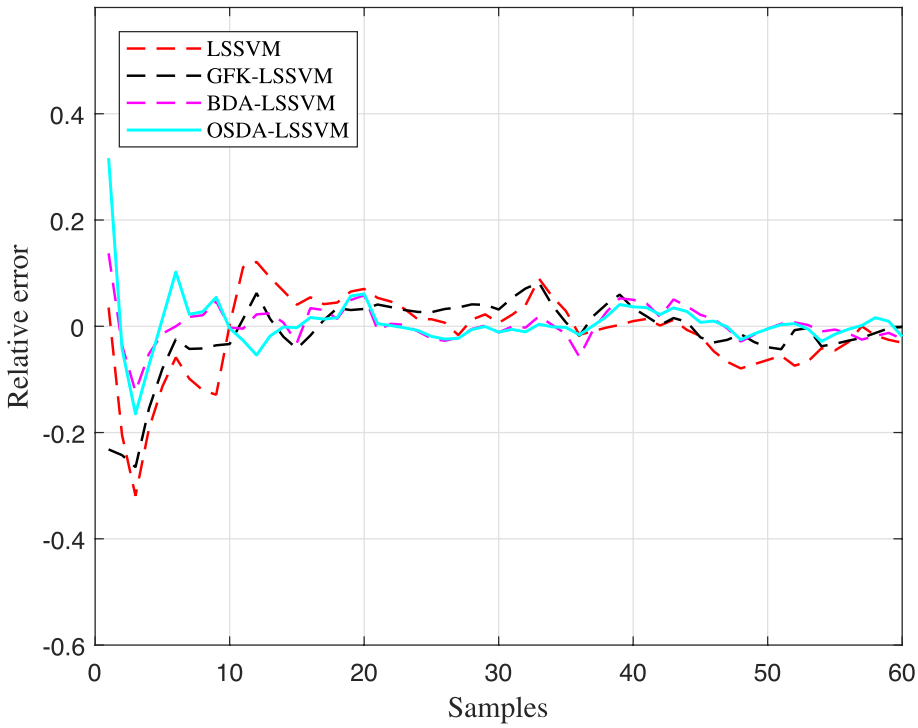

Fig. 11 Relative error curves of different soft sensor models in predicting inulinase concentration

Table 3 Assessment metrics for different models to predict *Pichia pastoris*

Soft sensor model	*Pichia pastoris* concentration		Inulinase concentration	
	RMSE	R^2	*RMSE*	R^2
LSSVM	1.8907	0.9749	0.1215	0.9832
GFK-LSSVM	1.5191	0.9838	0.0904	0.9907
BDA-LSSVM	1.1557	0.9906	0.0653	0.9952
OSDA-LSSVM	0.9695	0.9934	0.0548	0.9966

domains and approaches the two subspaces to reduce the data discrepancies, and proposes the OSDA method that utilizes the shared features of the two domains and the domain-specific features, which reduces the domain discrepancy in terms of both the statistic and geometrical structures. From Figs. 6, 7, 8 and 9, the overall prediction curve of the proposed OSDA-LSSVM model is able to fit the actual value and show good local performance. From Figs. 10, 11, 12, 13 and Table 3, the proposed model significantly reduces the model prediction error under multiple working conditions. The simulation results show that the OSDA-LSSVM soft sensor model based on the transfer learning strategy of substructure domain adaptation exhibits superior performance under multiple working conditions.

Of course, the OSDA-LSSVM model also has some limitations. Compared with online models, it is not able to update the model with newly generated samples in a timely manner, which may lead to model performance degradation. In addition, when the OSDA-LSSVM model is applied to different biochemical reaction processes, the number of primary and auxiliary variables needs to be determined manually, which is highly subjective. The changes of auxiliary variables in different biochemical reactions are significantly different, and the number of primary and auxiliary variables directly affects the response speed of the model. If the number of manually determined auxiliary variables is too high, it will increase the complexity of the model, and then affect the response speed of the model. On the contrary, if the number of auxiliary variables is too small, the complexity of the model will be reduced, resulting in the decline of prediction accuracy. In addition, since the fermentation process data set is obtained through offline sampling, the sample data is very limited, which limits the training effect of the soft sensor model to a certain extent.

In conclusion, it is necessary to further combine online learning in future research, and knowledge transfer in multi-source domains can solve the problem of sample limitation and the limitations of offline models. Meanwhile, the adaptive selection of primary and auxiliary variables can balance the response speed and prediction

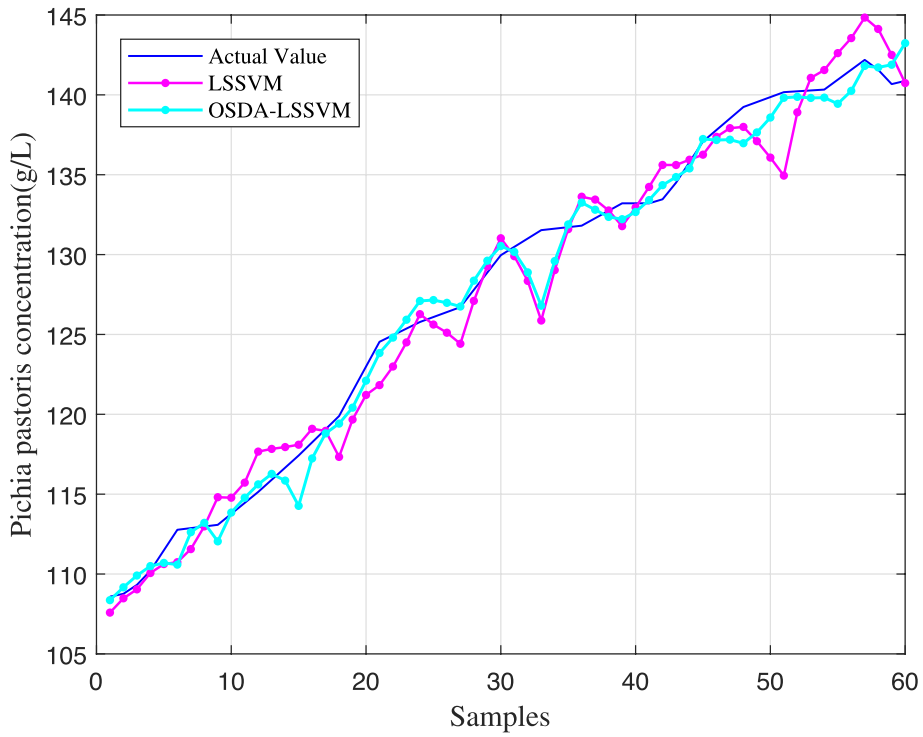

Fig. 12 The soft sensor model predicts *Pichia pastoris* concentration on the validation set

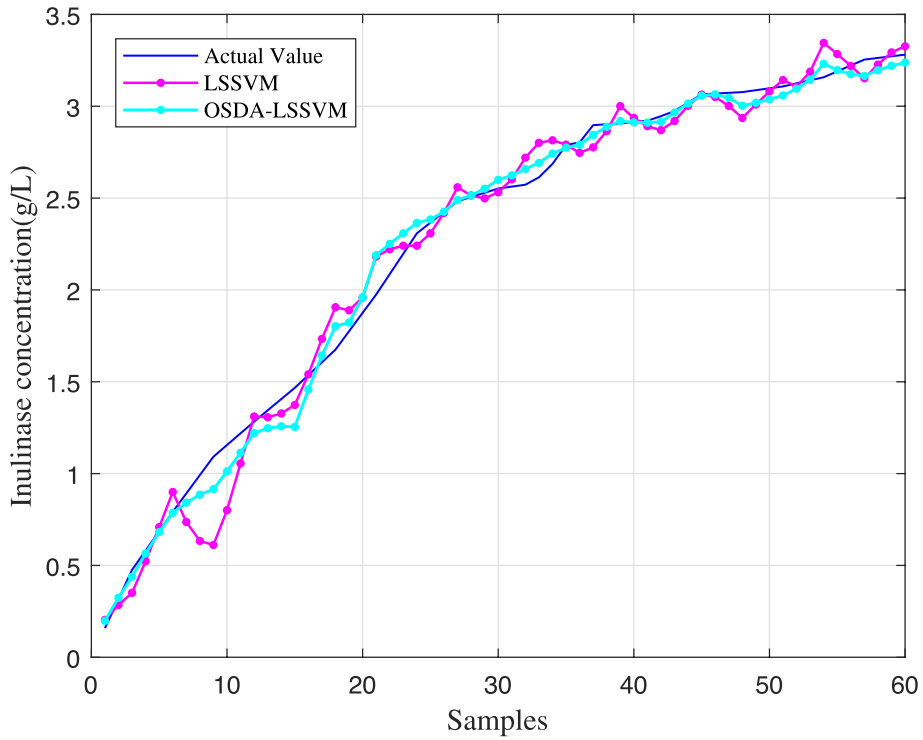

Fig. 13 The soft sensor model predicts inulinase concentration on the validation set

accuracy of the model to a certain extent, thus providing a basis for further model optimization and predictive control of biochemical reaction system.

Conclusion

The fermentation process of *Pichia pastoris* is characterized by multiple working condition, multiple stage and locality, and the performance of the traditional soft sensor model will be degraded or even model failure when the operating conditions are changed. In this paper, the OSDA-LSSVM soft sensor modeling method based on the modeling framework of substructure domain transfer is proposed. Aiming at the multi-stage characteristics of the fermentation process of *Pichia pastoris*, a transfer learning framework of substructure domain is introduced to carry out data transfer and modeling in different stages of *Pichia pastoris* fermentation, which effectively improves the local prediction performance of the model. Meanwhile, in order to solve the problem of data discrepancy caused by multiple working conditions, the OSDA-LSSVM soft sensor method combines data transfer and subspace alignment, utilizes multiple metric strategies (maximum variance, manifold regularization and distribution discrepancy minimization) to obtain the optimal projection matrices of the subspaces of the source and target domains, and projects the data of each sub-source domain and sub-target domain into the subspaces to perform subspace alignment. It reduces the data distribution discrepancy of different working conditions and retains the internal attributes and neighborhood structure of the original data, which effectively solves the model failure problem caused by multiple working conditions. Taking *Pichia pastoris* fermentation to produce inulinase as an example, different batches of data are used as source domain and target domain to verify the performance of the soft sensor model. The simulation results show that the OSDA -LSSVM model can accurately predict *Pichia pastoris* concentration and inulinase concentration online under different working conditions, which has higher prediction accuracy than the traditional soft sensor modeling method, and the method can be extended to other biological fermentation fields.

The model has advantages in terms of real-time and efficiency in the control of biochemical reactions, which is essential to optimize the performance and stability of the controller, making it highly suitable for industrial applications. In the field of process control, in order to improve the response efficiency of the system, the time complexity, memory algorithm complexity and computational complexity of the algorithm must be analyzed, but this is not the main focus of this paper, and we do not provide a detailed explanation. However, in the design and application of industrial biochemical

reaction control system, the above problems are worthy of further research and development.

Abbreviations

GMM	Gaussian mixture mode
BIC	Bayesian information criterion
OSDA	Optimal subspace domain adaptation
GFK	Geodetic flow kernel
PCA	Principal component analysis
BDA	Balanced distribution adaptation
MMD	Maximum mean discrepancy
MV	Maximum variance
MR	Manifold regularization
RBF	Radial basis function
FOS	Fructooligosaccharide
LSSVM	Least squares support vector machine
GFK-LSSVM	LSSVM predictive model based on Geodetic flow kernel
BDA-LSSVM	LSSVM predictive model based on balanced distribution adaptation
OSDA-LSSVM	LSSVM predictive model based on optimal subspace domain adaptation
RMSE	Root mean square error
R^2	Coefficient of determination

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12896-024-00928-4.

Supplementary Material 1.

Acknowledgements
We would like to acknowledge the hard and dedicated work of all the staff that implemented the intervention and evaluation components of the study.

Authors' contributions
Conceptualization, B.W. and J.W.; methodology, B.W.; software, J.W.; validation, J.W., L.Z. and C.J.; formal analysis, J.W.; investigation, J.W.; resources, J.W.; data curation, H.J.; writing-original draft preparation, J.W. and H.J.; writing-review and editing, S.H.; visualization, J.W.; supervision, B.W.; project administration, B.W.; funding acquisition, B.W. All authors have read and agreed to the published version of the manuscript.

Funding
This research was funded by the Natural Science Foundation of China (NO. 61705093), the Natural Science Foundation of the Jiangsu higher Education Institutions of China (NO.24KJA510011) and Wuxi "Light of Tai Lake" Science and Technology Project (basic research) (NO.K20221054).

Data availability
The data that support the findings of this study are available from Yangzhong Weikert Bioengineering Equipment Co., Ltd, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 7 September 2024 Accepted: 25 November 2024
Published online: 18 December 2024

References

1. Karbalaei M, Rezaee SA, Farsiani H. Pichia pastoris: A highly successful expression system for optimal synthesis of heterologous proteins. J Cell Physiol. 2020;235(9):5867–81.
2. Eskandari A, Nezhad NG, Leow TC, Rahman MBA, Oslan SN. Current achievements, strategies, obstacles, and overcoming the challenges of the protein engineering in Pichia pastoris expression system. World J Microbiol Biotechnol. 2023;40(1):39.
3. Mahboudi S, Shojaosadati SA, Maghsoudi A, Mahmoudi B. Development of a continuous fermentation process for the production of recombinant uricase enzyme by Pichia pastoris. Biotechnol Appl Biochem. 2024;71(1):123–31.
4. Zhao L, Li L, Hu M, Fang Y, Dong N, Shan A. Heterologous expression of the novel dimeric antimicrobial peptide LIG in Pichia pastoris. J Biotechnol. 2024;381:19–26.
5. Jyoti Gupta MS, Kumar Amit. Production of a Hepatitis E Vaccine Candidate Using the Pichia pastoris Expression System. Vaccine Des. 2022;2412:117–41.
6. Chai WY, Teo KTK, Tan MK, Tham HJ. Fermentation Process Control and Optimization. Chem Eng Technol. 2022;45(10):1731–47.
7. Wang B, Wang X, He M, Zhu X. Study on Multi-Model Soft Sensor Modeling Method and Its Model Optimization for the Fermentation Process of Pichia pastoris. Sensors. 2021;21(22):7635.
8. Sun Ym, Du N, Sun Qy, Chen Xg, Yang Jw. Research and application of biological potency soft sensor modeling method in the industrial fed-batch chlortetracycline fermentation process. Clust Comput. 2019;22(Suppl 3):S6019–S6030.
9. Qiu K, Wang J, Zhou X, Wang R, Guo Y. Soft sensor based on localized semi-supervised relevance vector machine for penicillin fermentation process with asymmetric data. Measurement. 2022;202: 111823.
10. Hua L, Zhang C, Sun W, Li Y, Xiong J, Nazir MS. An evolutionary deep learning soft sensor model based on random forest feature selection technique for penicillin fermentation process. ISA Trans. 2023;136:139–51.
11. Dave N, Varadavenkatesan T, Selvaraj R, Vinayagam R. Modelling of fermentative bioethanol production from indigenous Ulva prolifera biomass by Saccharomyces cerevisiae NFCCI1248 using an integrated ANN-GA approach. Sci Total Environ. 2021;791: 148429.
12. Yamada N, Kaneko H. Adaptive soft sensor ensemble for selecting both process variables and dynamics for multiple process states. Chemom Intell Lab Syst. 2021;219: 104443.
13. Chai Z, Zhao C, Huang B, Chen H. A Deep Probabilistic Transfer Learning Framework for Soft Sensor Modeling With Missing Data. IEEE Trans Neural Netw Learn Syst. 2022;33(12):7598–609.
14. Xie J, Huang B, Dubljevic S. Transfer Learning for Dynamic Feature Extraction Using Variational Bayesian Inference. IEEE Trans Knowl Data Eng. 2022;34(11):5524–35.
15. Ren JC, Liu D, Wan Y. VMD-SEAE-TL-Based Data-Driven soft sensor modeling for a complex industrial batch processes. Measurement. 2022;198: 111439.
16. Zhou X, Sbarufatti C. A fuzzy-set-based joint distribution adaptation method for regression and its application to online damage quantification for structural digital twin. Mech Syst Signal Process. 2023;191: 110164.
17. Liu Y, Yang C, Zhang M, Dai Y, Yao Y. Development of Adversarial Transfer Learning Soft Sensor for Multigrade Processes. Ind Eng Chem Res. 2020;59(37):16330–45. https://doi.org/10.1021/acs.iecr.0c02398.
18. Zhu J, Dai Y, Guo W, Deng H, Liu Y. Domain Compensation-Assisted Quality Inference Enhancement of Chemical Processes with Distributed Outputs. Ind Eng Chem Res. 2024;63(8):3632–40. https://doi.org/10.1021/acs.iecr.3c04480.
19. Liu Y, Yang C, Liu K, Chen B, Yao Y. Domain adaptation transfer learning soft sensor for product quality prediction. Chemom Intell Lab Syst. 2019;192: 103813. https://doi.org/10.1016/j.chemolab.2019.103813.
20. Lu W, Chen Y, Wang J, Qin X. Cross-domain activity recognition via substructural optimal transport. Neurocomputing. 2021;454:65–75.

21. Zhao J, Deng F, He H, Chen J. Local Domain Adaptation for Cross-Domain Activity Recognition. IEEE Trans Hum Mach Syst. 2021;51(1):12–21.
22. Wang Z, Wang X, Liu F, Gao P, Ni Y. Adaptative Balanced Distribution for Domain Adaptation with Strong Alignment. IEEE Access. 2021;9:100665–76.
23. Wu D, Lawhern V, Gordon S, Lance B, Lin C. Driver Drowsiness Estimation from EEG Signals Using Online Weighted Adaptation Regularization for Regression (OwARR)(Article). IEEE Trans Fuzzy Syst. 2017;25(6):1522–35.
24. Gholenji E, Tahmoresnezhad J. Joint discriminative subspace and distribution adaptation for unsupervised domain adaptation. Appl Intell. 2020;50(7):2050–66.
25. Xing Z, Peng J, He X, Tian M. Semi-supervised sparse subspace clustering with manifold regularization. Appl Intell. 2024;54(9):6836–45.
26. Belkin M, Niyogi P, Sindhwani V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. J Mach Learn Res. 2006;7:2399–434.
27. Suykens JAK, Vandewalle J. Least Squares Support Vector Machine Classifiers. Neural Process Lett. 1999;9(3):293–300.

Publisher's Note