Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Ensemble enhanced active learning mixture discriminant analysis model and its application for semi-supervised fault classification*#

Weijun WANG[1], Yun WANG[2], Jun WANG[1], Xinyun FANG[3], Yuchen HE‡[1]

[1]*Key Laboratory of Intelligent Manufacturing Quality Big Data Tracing and Analysis of Zhejiang Province, China Jiliang University, Hangzhou 310018, China*

[2]*Mechanical and Electrical Engineering Department, Zhejiang Tongji Vocational College of Science and Technology, Hangzhou 311231, China*

[3]*Suzhou Institute of Metrology, Suzhou 215004, China*

E-mail: S20010811027@cjlu.edu.cn; wy@zjtongji.edu.cn; blackknight@cjlu.edu.cn; fangxy@szjl.com.cn; yche@cjlu.edu.cn

Received Feb. 13, 2022; Revision accepted May 9, 2022; Crosschecked May 20, 2022; Published online July 26, 2022

Abstract: As an indispensable part of process monitoring, the performance of fault classification relies heavily on the sufficiency of process knowledge. However, data labels are always difficult to acquire because of the limited sampling condition or expensive laboratory analysis, which may lead to deterioration of classification performance. To handle this dilemma, a new semi-supervised fault classification strategy is performed in which enhanced active learning is employed to evaluate the value of each unlabeled sample with respect to a specific labeled dataset. Unlabeled samples with large values will serve as supplementary information for the training dataset. In addition, we introduce several reasonable indexes and criteria, and thus human labeling interference is greatly reduced. Finally, the fault classification effectiveness of the proposed method is evaluated using a numerical example and the Tennessee Eastman process.

Key words: Semi-supervised; Active learning; Ensemble learning; Mixture discriminant analysis; Fault classification

https://doi.org/10.1631/FITEE.2200053 **CLC number:** TP277

1 Introduction

In the past decades, to ensure process safety and improve product quality in modern industries, many process monitoring methods have been proposed; among them, detection and fault classification play significant roles in guaranteeing the effective operation of process monitoring systems (Chiang et al., 2004; Ge et al., 2013; Deng et al., 2022). Fault classification techniques are generally employed to determine the type of fault that has occurred by analyzing the discrepancy and similarity between faults. Meanwhile, with the rapid development of information technology, such as distributed control systems

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (No. 61903352), the Natural Science Foundation of Zhejiang Province, China (No. LQ19F030007), the Project of Department of Education of Zhejiang Province, China (No. Y202044960), the China Postdoctoral Science Foundation (No. 2020M671721), the Fundamental Research Funds for the Provincial Universities of Zhejiang, China (Nos. 2021YW18, 2021YW80, and 2022YW96), and the Innovative Team Project of Fujian Institute of Metrology, China

Electronic supplementary materials: The online version of this article (https://doi.org/10.1631/FITEE.2200053) contains supplementary materials, which are available to authorized users

ORCID: Weijun WANG, https://orcid.org/0000-0003-3655-4569; Yun WANG, https://orcid.org/0000-0002-7512-0168; Jun WANG, https://orcid.org/0000-0002-2742-3041; Yuchen HE, https://orcid.org/0000-0002-0528-2778

(DCSs) and the Internet of Things (IoT), massive valuable data have been collected, transmitted, and stored. Consequently, data-driven methods have recently become feasible methods for fault classification (MacGregor and Cinar, 2012; Ge, 2017, 2018; Ge et al., 2017).

As typical data-driven approaches, traditional multivariate statistical process monitoring (MSPM) methods have been widely used in process monitoring and achieved great success in recent years (Feng et al., 2013; Ge, 2016). For example, to extract potential variables, principal component analysis (PCA) is proposed as an effective dimension reduction method for fault detection (Dong and Qin, 2018), while partial least squares (PLSs) have better applicability for the common input-output modeling in chemical processes (Botre et al., 2017; Harkat et al., 2019). Nevertheless, correlation among different modes is always neglected during characteristic extraction, which makes these methods unsuitable for fault classification. Alternatively, Fisher discriminant analysis (FDA), which finds appropriate decision surfaces among different classes by minimizing within-class variance and maximizing between-class diversity, has been considered as one of the most effective solutions (Chiang et al., 2000; Zhong et al., 2020).

However, the classification performance of FDA relies heavily on the assumption that data are Gaussian distributed. This assumption means that fault types cannot be identified in real situations, because practical process data are often non-Gaussian distributed due to the complexity of the system setting (Abramson et al., 1963). Recently, to handle this problem, traditional linear discriminant analysis was extended to a mixed form, known as the mixture discriminant analysis (MDA) model, in which every class can be estimated by a Gaussian mixture model (GMM) (Hastie and Tibshirani, 1996; Fraley and Raftery, 2002). Compared with FDA, MDA has excellent classification performance when dealing with within-class non-Gaussianity. For example, Kalantar et al. (2019) evaluated the performance of FDA and MDA in groundwater potential mapping. The results showed that MDA has better robustness and applicability than FDA. Huang et al. (2013) proposed a hybrid multivariate approach named mixture discriminant monitoring (MDM), which overcomes the defects of FDA and identifies the unknown-fault-related variables using statistical process control (SPC) chart techniques. In addition, Pu and Li (2021) proposed a probability discrimination algorithm combined with information theoretic learning (ITL) and robust MDA (RMDA) to address the industrial label-noise fault diagnosis issue.

Note that the MDA method usually requires adequate sample labels, which cannot be easily obtained in real situations because the accumulation of expert experiences and prior process knowledge are often time-consuming and costly (Wang J et al., 2014). Therefore, the labeled samples in an actual industry are often very sparse, and insufficient labeling for supervised modeling will lead to inadequate statistical significance of the model and cause overfitting which degrades the classification performance. Conversely, unlabeled samples are often abundant and can reflect a great deal of information. As a result, a semi-supervised fault classification strategy has become an increasingly hot topic and has attracted much attention recently. A large number of semi-supervised learning (SSL) methods (Schwenker and Trentin, 2014; Liu JW et al., 2015) have been proposed to acquire the information in unlabeled data, including the self-training model (Raina et al., 2007; Zou et al., 2019), semi-supervised support vector machines (SVMs) (Chapelle et al., 2006; Wang L et al., 2021), graph-based SSL (Blum and Chawla, 2001; Chen et al., 2018; Zaman and Liang, 2021), committee-based SSL (Hady and Schwenker, 2010; Cui et al., 2012), and semi-supervised generative model (Yan et al., 2014; Yao and Ge, 2017). Among them, the self-training model adds non-informative samples to the dataset to improve the performance of the classifier. Semi-supervised SVMs can use unlabeled samples to optimize decision boundaries and place those boundaries in the low-density area, whereas graph-based SSL methods can transfer the relationship between samples into a graph and can label the unlabeled samples using the min cut algorithm. Committee-based SSL methods try to construct a variety of classifiers that can cooperatively use the information in unlabeled samples. Compared with other semi-supervised methods, all available samples share the same model in the semi-supervised generative model, which has received much attention in the past decades (Yan et al., 2014; Yao and Ge, 2017). For example, Yan et al. (2014) extended the MDA model to a semi-supervised MDA (SMDA)

structure, which is more conducive to dealing with non-Gaussian distributed data with missing labels. Farajzadeh-Zanjani et al. (2021) proposed a semi-supervised framework incorporating generative adversarial networks (GANs) to generate minority class samples. Particularly, soft sensing frameworks for the insufficient labeling issue have been discussed and corresponding solutions (such as mixture factor analysis and mixtures of extreme learning machines) have been proposed (Shao and Tian, 2017; Shao et al., 2019a, 2019b; Wang JB et al., 2019).

However, the information of labeled samples is still limited and cannot fully describe the behavior of unlabeled samples. The performance of semi-supervised methods will severely deteriorate as unlabeled samples accumulate. Some high-value unlabeled samples are crucial in dealing with the classification issue, which should be labeled and chosen for model construction. Alternatively, the active learning technique has drawn much attention recently. This technique efficiently combines process knowledge and expert experience, measures the value of each unlabeled sample, and considers it in semi-supervised modeling using human intervention. The main idea of active learning is to collect suitable samples from an unlabeled dataset under the guidance of supervised data, to provide supplementary process information to the original labeled dataset (Settles, 2012). Active learning can add the most relevant unlabeled samples to the training dataset by setting appropriate indexes that measure the value of each sample, and it fully considers both labeled and unlabeled information through a continuous interactive learning process. For example, Liu J et al. (2018) proposed a semi-supervised exponential discriminant analysis method based on active learning (AL-SEDA), which not only extends the traditional EDA to the fault classification application in semi-supervised industrial processes, but also alleviates the noise introduced by the random labeling of unlabeled samples. On this basis, an active learning based semi-supervised FDA (ALSemiFDA) was created, in which entropy was used to judge the value of unlabeled samples and then employed for sample labeling (Yin et al., 2018, 2019). Recently, active learning methods have been explored for actual bearing compound fault diagnosis; the results showed that such methods can achieve high diagnostic accuracy with only a small number of high-value samples, which greatly reduces the labeling workload of domain experts (Jin et al., 2021). However, it should be mentioned that the selection of unlabeled samples depends heavily on one single index and the original model derived from the insufficient labeled training dataset, which may lead to several severe problems. First, one single index, such as entropy, is inappropriate and inadequate to judge the value of the unlabeled sample. As a result, other related indexes should be involved to help strengthen the performance of unlabeled sample evaluation. Second, human labeling is always costly and time-consuming in real situations, whereas model labeling can solve the dilemma and improve the efficiency of sample labeling. Finally, one unified model based on an insufficient labeled training dataset may result in the overfitting problem and can be easily affected by the wrongly unlabeled samples.

To solve the above problems, an ensemble enhanced MDA based on active learning (called E^{2}ALMDA) is proposed, where ensemble learning is adopted to guarantee the reliability of fault classification results. The ensemble learning method naturally improves the robustness of semi-supervised methods by integrating the classification results of multiple weak sub-classifiers formed by different randomly collected training sub-datasets under formulated rules (Araya et al., 2017; Liu Y and Ge, 2018; Zheng et al., 2019; He et al., 2021; Zhang et al., 2022). In each sub-classifier, active learning is then applied to introduce suitable unlabeled samples into the corresponding training sub-dataset. In this way, active learning extracts the most related information contained in the unlabeled data and uses it to solve semi-supervised problems. Compared with previous works, the value of each unlabeled sample is evaluated according to entropy, prediction error, and two newly designed indexes (confidence and deficiency). In this way, human labeling in traditional active learning methods is replaced by model labeling, which can effectively avoid human intervention. Finally, the K-nearest neighbor (KNN) method is employed in the ensemble learning framework to obtain the final classification results by integrating the performance of each sub-classifier based on the posterior probability matrices of labeled samples and test samples on all possible sub-classifiers. To verify the effectiveness of the proposed method, two cases, including a numerical example and the

Tennessee Eastman process (TEP) benchmark, are adopted where the fault classification results of different methods are discussed in detail.

The main contributions of this paper can be summarized as follows:

1. Several selection indexes are proposed to evaluate the performance of unlabeled samples in various aspects.

2. Human labeling in active learning is replaced by model labeling; the reliability of model labeling is improved according to the four proposed indexes, while human interference is also avoided.

3. A reasonable stopping criterion is proposed to introduce more informative samples to the sub-dataset.

The notations used throughout the paper are listed in Table 1.

2 ALSemiFDA

As an integration of active learning and the SemiFDA method, ALSemiFDA tackles the semi-supervised problem based on all available data (Yin et al., 2018, 2019). On one hand, the active learning in ALSemiFDA can improve the classification performance by manually introducing valuable unlabeled samples into a labeled dataset according to expert query. On the other hand, SemiFDA tries to make use of the unlabeled dataset to provide supplementary information for the traditional FDA method. The detailed learning steps of ALSemiFDA can be expanded as follows:

1. The traditional FDA model is initialized according to the labeled dataset.

2. The unlabeled data are labeled by the initial FDA model and then chosen as new members of the labeled dataset.

3. The entropy e_i of each unlabeled sample x_i is obtained based on the classification results from the SemiFDA model. The sample x_{s} with the maximum entropy will be added to the training labeled dataset. Hence, e_i and x_{s} can be defined as follows:

$$e_i = -\sum_{j=1}^{J} \left(p_{ij} \ln p_{ij} \right), \tag{1}$$

$$x_{\mathrm{s}} = \arg \max_{x_i \in X_{\mathrm{U}}} e_i. \tag{2}$$

4. Steps 2 and 3 are executed repeatedly until the maximal classification probability of all unlabeled samples with respect to any class reaches the threshold t, which can be expressed as follows:

$$\max_{1 \le j \le J} p_{ij} \ge t, \ \forall x_i \in X_{\mathrm{U}}. \tag{3}$$

Although the ALSemiFDA model takes unlabeled samples into consideration and the corresponding classification results are enhanced to some extent, the following issues should be covered:

1. The single criterion based on entropy is quite fragile and the classification results can be easily

Table 1 Nomenclature

Symbol	Description	Symbol	Description
X_{L}	Labeled dataset	t	Threshold
X_{U}	Unlabeled dataset	x_{new}	The new observation sample
m	Number of variables	C^g	Confusion matrix of the g^{th} sub-dataset
n	Number of labeled samples	Confidence($i|g$)	Confidence of the i^{th} unlabeled sample in the g^{th} sub-classifier
G	Number of sub-classifiers		
J	Number of classes	Deficiency($i|g$)	Deficiency of the i^{th} unlabeled sample in the g^{th} sub-classifier
K_j	Number of subclasses in the j^{th} class		
c_{jk}	The k^{th} subclass in the j^{th} class	Error($i|g$)	Prediction error of the i^{th} unlabeled sample in the g^{th} sub-classifier
Π_j	Priori probability of the j^{th} class		
π_{jk}	Priori probability of k^{th} subclass in the j^{th} class	Value($i|g$)	Value of the i^{th} unlabeled sample in the g^{th} sub-classifier
μ_{jk}	Mean of the k^{th} subclass in the j^{th} class	P_x	Posterior probability matrix of the x^{th} labeled sample
Σ_{jk}	Variance of the k^{th} subclass in the j^{th} class		
x_i	The i^{th} sample	PL	The series of posterior probability matrices of all labeled sample
θ	Parameter set of MDA		
e_i	Entropy of x_i	P_{new}	Posterior probability matrix of x_{new}
p_{ij}	Posteriori probability of x_i to class j	D_{ix}	Euclidean distance between P_{new} and P_x
x_{s}	Selected unlabeled sample		

MDA: mixture discriminant analysis

affected by the disturbance caused by the inappropriate involvement of unlabeled samples.

2. The uniform model established based on all labeled samples can probably be overfitted and influence the introduction of unlabeled samples into the labeled dataset.

3. Human labeling generally requires plenty of human intelligence, which can be costly and time-consuming. It is increasingly difficult to rely on expert knowledge due to the data dimension explosion, while insufficient expert query may result in severe misclassification.

4. The basic assumption of ALSemiFDA is that data are Gaussian distributed, which cannot be easily met in real situations due to non-Gaussianity. Meanwhile, similar to the traditional FDA model, the classification results can be influenced by noise.

3 E^{2}ALMDA

E^{2}ALMDA is proposed and discussed in detail in this section to cope with the above problems. First, the bagging technique is introduced to randomly collect labeled samples to construct sub-classifiers for subsequent integration. Then, traditional active learning will be enhanced by several newly designed indexes to help improve the reliability of the involved unlabeled samples. Finally, ensemble learning will be used to strengthen the robustness of the classification results by integrating the performance of all sub-classifiers.

3.1 Weak classifier construction based on the MDA model

The bagging technique is used in our work to obtain several sub-datasets by random sampling (Dietterich, 2000; Abellán and Masegosa, 2010). Assume that the labeled dataset is $X_{\mathrm{L}} = [x_1, x_2, ..., x_n] \in \mathbb{R}^{m\times n}$, where x_i $(i = 1, 2, ..., n)$ denotes the corresponding sample, and m and n represent the numbers of variables and samples, respectively. Then, G different training datasets can be derived based on the bagging technique, which can form corresponding sub-classifiers. In the g^{th} $(g = 1, 2, ..., G)$ sub-dataset, due to a sophisticated process, the data collections are unlikely to follow Gaussian distributions. Therefore, the MDA technique is adopted where the data distribution is regarded as a mixture of multiple Gaussian components, to better fit the non-Gaussian characteristics of each sub-dataset (Fraley and Raftery, 2002; Bouveyron and Girard, 2009). In the MDA method, it is assumed that all training data consist of J classes, where each class j $(j = 1, 2, ..., J)$ can be further divided into K_j artificial subclasses, denoted by c_{jk} $(k = 1, 2, ..., K_j)$ which obeys the Gaussian distribution $N(\mu_{jk}, \Sigma_{jk})$. The prior probability that one sample belongs to class j can be denoted by Π_j, and the mixing probability of the k^{th} subclass within class j is represented by the parameter π_{jk} where $\sum_{j=1}^{J}\Pi_j = 1$, $\sum_{k=1}^{K_j}\pi_{jk} = 1$.

Therefore, the mixture density for class j can be expressed as follows (Yan et al., 2014):

$$\begin{aligned} h_j(x) =& p(x|j) \\ =& \sum_{k=1}^{K_j}\left\{\pi_{jk}(2\pi)^{-m/2}|\Sigma_{jk}|^{-1/2}\right. \\ & \left.\cdot\exp\left[-\frac{(x-\mu_{jk})^{\mathrm{T}}\Sigma_{jk}^{-1}(x-\mu_{jk})}{2}\right]\right\}. \end{aligned} \tag{4}$$

The parameter set $\theta = \{\mu_{jk}, \Sigma_{jk}, \pi_{jk}\}$ $(j = 1, 2, ..., J,\ k = 1, 2, ..., K_j)$ can be estimated by maximizing the following log-likelihood function:

$$\theta = \arg\max_{\theta} l(\theta) = \arg\max_{\theta}\sum_{i=1}^{n}\ln p(x_i|\theta). \tag{5}$$

The above problem can be solved by the expectation-maximization (EM) algorithm. In the expectation step (E-step), provided that x_i is a training sample in class j, the probability of x_i belonging to the k^{th} subclass in class j can be estimated as Eq. (6) (on the top of the next page), which is followed by the maximization step (M-step):

$$\pi_{jk} = \frac{\sum_{x_i\in j} p(c_{jk}|x_i, j)}{\sum_{q=1}^{K_j}\sum_{x_i\in j} p(c_{jq}|x_i, j)}, \tag{7}$$

$$\mu_{jk} = \frac{\sum_{x_i\in j}\left(x_i p(c_{jk}|x_i, j)\right)}{\sum_{x_i\in j} p(c_{jk}|x_i, j)}, \tag{8}$$

$$\Sigma_{jk} = \frac{\sum_{x_i\in j}\left(p(c_{jk}|x_i, j)(x_i-\mu_{jk})(x_i-\mu_{jk})^{\mathrm{T}}\right)}{\sum_{x_i\in j} p(c_{jk}|x_i, j)}. \tag{9}$$

The above two steps are repeated iteratively until convergence. Then, the posterior probabilities of a new observation x_{new} belonging to class j can be

$$p(c_{jk}|x_i,j)=\frac{p(x_i,c_{jk}|j)}{p(x_i|j)}=\frac{p(c_{jk}|j)p(x_i|c_{jk},j)}{p(x_i|j)}$$
$$=\frac{\pi_{jk}(2\pi)^{-m/2}|\Sigma_{jk}|^{-1/2}\exp\left[-\frac{(x_i-\mu_{jk})^{\mathrm{T}}\Sigma_{jk}^{-1}(x_i-\mu_{jk})}{2}\right]}{\sum_{q=1}^{K_j}\left\{\pi_{jq}(2\pi)^{-m/2}|\Sigma_{jq}|^{-1/2}\exp\left[-\frac{(x_i-\mu_{jq})^{\mathrm{T}}\Sigma_{jq}^{-1}(x_i-\mu_{jq})}{2}\right]\right\}}. \tag{6}$$

computed using the Bayesian theorem:

$$\begin{aligned}&p(j|x_{\mathrm{new}})\\=&\frac{p(x_{\mathrm{new}}|j)p(j)}{p(x_{\mathrm{new}})}\propto p(j)p(x_{\mathrm{new}}|j)\\=&\Pi_j\sum_{k=1}^{K_j}\left\{\pi_{jk}(2\pi)^{-m/2}|\Sigma_{jk}|^{-1/2}\right.\\&\left.\cdot\exp\left[-\frac{(x_{\mathrm{new}}-\mu_{jk})^{\mathrm{T}}\Sigma_{jk}^{-1}(x_{\mathrm{new}}-\mu_{jk})}{2}\right]\right\}.\end{aligned} \tag{10}$$

The new observation x_{new} can be assigned to a specific class using the maximum a posteriori (MAP) rule, that is, $j_{\mathrm{new}}=\arg\max_j p(j|x_{\mathrm{new}})$.

3.2 Enhanced active learning for weak classifiers

As an effective solution for extracting the inherent supplementary information within the unlabeled dataset, traditional active learning methods usually set different indexes, such as entropy or prediction error, to evaluate the importance of each unlabeled sample. However, the introduction of unlabeled samples in the labeled dataset is still based on human labeling, which seems subjective and time-consuming (Snow et al., 2008; Ipeirotis et al., 2010; Raykar et al., 2010; Yuen et al., 2011). Actually, model labeling can significantly avoid these situations, and the only problem is to ensure the modeling reliability. Unfortunately, entropy and prediction error seem to be inadequate to realize the precision requirement.

On this basis, two new selection indexes, called confidence and deficiency, are designed in this study to realize model labeling for unlabeled samples. First, a verification dataset is randomly collected from the labeled samples. Then the following confusion matrix of the g^{th} sub-dataset can be obtained based on the classification result of the verification dataset using the g^{th} sub-classifier:

$$C^g=\begin{pmatrix}N_{11}^g & N_{12}^g & \cdots & N_{1J}^g\\ N_{21}^g & N_{22}^g & \cdots & N_{2J}^g\\ \vdots & \vdots & \vdots & \vdots\\ N_{J1}^g & N_{J2}^g & \cdots & N_{JJ}^g\end{pmatrix}, \tag{11}$$

where N_{aj}^g $(a=1,2,\ldots,J)$ represents the number of samples in class a being misclassified into class j by the g^{th} sub-classifier. Meanwhile, two important variables are defined here, $c_j^g=N_{jj}^g/\sum_{a=1}^J N_{aj}^g$ and $r_j^g=1-N_{jj}^g/\sum_{a=1}^J N_{ja}^g$. The first variable represents the credibility that one sample is classified to the j^{th} class using the g^{th} sub-classifier, and the second variable indicates the degree to which the g^{th} sub-classifier lacks the information of class j.

The confidence and deficiency indexes can be defined as follows:

$$\mathrm{Confidence}(i|g)=\sum_{j=1}^{J}\left(p_{ij}^g c_j^g\right), \tag{12}$$

$$\mathrm{Deficiency}(i|g)=\sum_{j=1}^{J}\left(p_{ij}^g r_j^g\right), \tag{13}$$

where p_{ij}^g indicates the probability of the i^{th} unlabeled sample belonging to class j using the g^{th} sub-classifier. The confidence index (Eq. (12)) indicates the reliability of fault classification with respect to the i^{th} unlabeled sample using the g^{th} sub-classifier. Unlabeled samples with high confidence are likely to be added to the labeled dataset of the corresponding sub-classifier. On the contrary, a high value of the deficiency index indicates that the g^{th} classifier lacks information of the i^{th} unlabeled sample, which seems to be the perfect supplementary information to the current sub-classifier. In this way, four indexes, entropy, prediction error, confidence, and deficiency, are integrated to provide a comprehensive

explanation of model labeling:

$$e_i^g = -\sum_{j=1}^{J} \left(p_{ij}^g \ln p_{ij}^g \right), \tag{14}$$

$$\text{Error}(i|g) = \frac{\sum_{m=1}^{n_\text{U}} \left[1 - p_{i,g}^{(\theta+1)}(\hat{y}_m) \right]}{n_\text{U}}, \tag{15}$$

where e_i^g indicates the entropy of the i^{th} unlabeled sample using the g^{th} sub-classifier. Meanwhile, $p_{i,g}^{(\theta+1)}(\hat{y}_m)$ represents the probability that the m^{th} unlabeled sample is identified as the g^{th} subclassifer in which the i^{th} unlabeled sample has been added previously. Therefore, $\text{Error}(i|g)$ is the average classification performance of the g^{th} sub-classifier after it is enhanced by adding the i^{th} unlabeled sample. Finally, the value of the i^{th} unlabeled sample with respect to the g^{th} sub-classifier can be defined as follows:

$$\begin{cases} \text{Value}(i|g) = e_i^g + \text{Confidence}(i|g) \\ \qquad\qquad\quad +\text{Deficiency}(i|g) - \text{Error}(i|g), \\ x_\text{s}^g = \arg \max\limits_{x_i \in X_\text{U}} \text{Value}(i|g). \end{cases} \tag{16}$$

Note that each index should be normalized in the value estimation. It can be seen that entropy, confidence, and deficiency are positively correlated with the final value, while the prediction error is negatively correlated. This means that unlabeled samples with high entropy, high confidence, high deficiency, and low prediction error are likely to be considered as supplementary information for the g^{th} sub-classifier. The unlabeled sample with the highest "Value" is selected as the candidate to join the current g^{th} sub-dataset, which will be updated iteratively until the following criterion is satisfied:

$$\frac{\sum_{i=1}^{n_\text{U}^{(\theta+1)}} \text{Var}^{(\theta+1)}(p_i|g)}{n_\text{U}^{(\theta+1)}} < \frac{\sum_{i=1}^{n_\text{U}^{(\theta)}} \text{Var}^{(\theta)}(p_i|g)}{n_\text{U}^{(\theta)}}, \tag{17}$$

where $\text{Var}^{(\theta)}(p_i|g)$ indicates the variance of the classification results of the i^{th} unlabeled sample using the g^{th} sub-classifier. The above restriction guarantees the continuous enhancement of each sub-classifier during the iteration process. That is why the proposed method is called enhanced active learning.

3.3 Information fusion for fault classification

After each sub-classifier is strengthened, the results of different sub-classifiers should be integrated. Here, the verification dataset is employed again to obtain the following posterior probability matrix:

$$P_x = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1J} \\ p_{21} & p_{22} & \cdots & p_{2J} \\ \vdots & \vdots & \vdots & \vdots \\ p_{G1} & p_{G2} & \cdots & p_{GJ} \end{pmatrix}, \tag{18}$$

where p_{qj} $(q = 1, 2, \ldots, G)$ represents the posterior probability that the current sample is identified as a member of the j^{th} class using the q^{th} subclassifier. Therefore, a series of posterior probability matrices can be derived for these labeled samples, $\text{PL} = \{P_1, P_2, ..., P_n\}$, which are further considered as the reference of final classification results.

Assume that the posterior probability matrix of test sample x_new is denoted by P_new. Then the Euclidean distance D_{ix} between P_new and the posterior probability matrix of each labeled sample P_x can be expressed as follows:

$$D_{ix} = \|P_\text{new} - P_x\|_\text{F}, \tag{19}$$

where "F" denotes the Frobenius norm.

According to the above formula, we can find K labeled samples closest to the sample x_new and count the number belonging to class j among these K samples, expressed as n_j. The class with the largest n_j is chosen as the final classification result, which can be written as follows:

$$\text{Final}(i) = \arg \max_j n_j, \ j = 1, 2, ..., J. \tag{20}$$

The detailed E^{2}ALMDA flowchart is shown in Fig. 1, and the whole procedure is expressed in Algorithm 1. Fig. 1 shows that E^{2}ALMDA can be disassembled into three parts. First, ensemble learning can extract diverse information from sparse labeled data and construct a unique validation dataset as the basis for calculating confidence and deficiency. Second, enhanced active learning adds more high-value samples to each sub-dataset, to improve the performance of each sub-classifier. Finally, data fusion takes into account the diversity of all sub-classifiers to comprehensively predict the labels of the test samples.

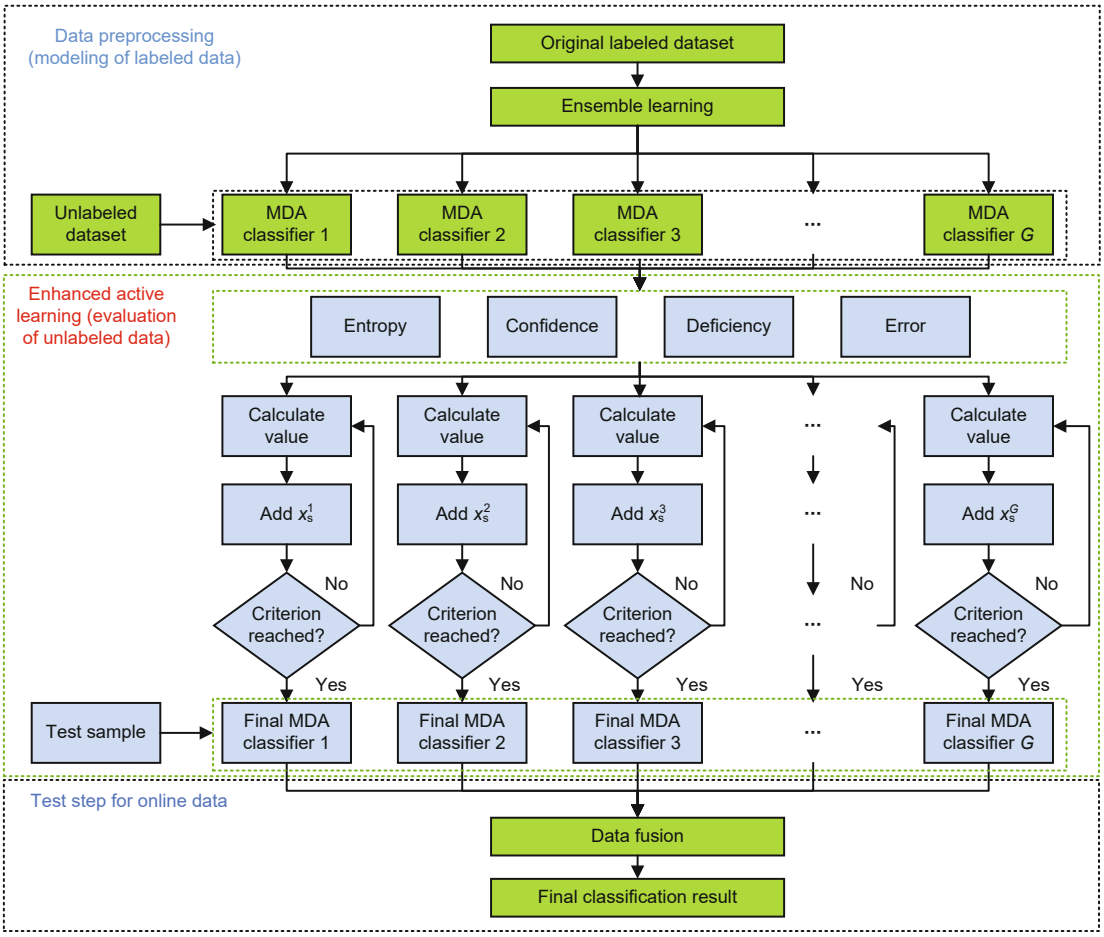

Fig. 1 Flowchart of E²ALMDA modeling and testing

4 Case study

In this section, a numerical example and the TEP are introduced to evaluate the performance of the proposed method. In addition, several related methods, including MDA, SMDA, ALMDA, and ALSemiFDA, are considered for comparison. Note that ALMDA usually treats entropy as the selection index and overall prediction error as the stop criterion, which is entirely distinct from the proposed E²ALMDA.

4.1 Numerical example

The purpose of this experiment is to discriminate a dataset consisting of five types of non-Gaussian data, each of which contains two variables. A training dataset of 200 samples and a testing dataset of 100 samples are collected. The specific configuration of the dataset is shown in Table 2. To simulate insufficient labeling, only 20% of the generated training samples can keep corresponding labels, while the others are considered as unlabeled samples. The accuracy results of various methods are shown in Table 3.

In Table 3, it can be seen that the five methods perform quite differently due to the scarcity of labeled samples. Testing samples can be easily misclassified using MDA, especially in classes 1, 4, and 5, because only the labeled samples are involved in modeling, while abundant unlabeled samples are directly abandoned, which results in significant missing data. Although SMDA performs better than MDA because unlabeled samples are introduced, it still has some difficulties in identifying specific types of data, such as class 4. The main reason is that the labels of unlabeled samples are assigned by the supervised model, which is constructed from labeled samples. As a result, it is quite unreasonable that the labeling

Table 2 Configuration of the numerical example

Class	Number of components	Component proportion	Mean	Variance	Number of samples for training	Number of samples for testing
1	5	(0.1, 0.1, 0.6, 0.1, 0.1)	$\begin{bmatrix} -2 & -1 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	$0.3I_2$	200*	100
2	5	(0.1, 0.1, 0.6, 0.1, 0.1)	$\begin{bmatrix} -2 & -1 & 0 & 1 & 2 \\ 3 & 3 & 3 & 3 & 3 \end{bmatrix}$	$0.3I_2$	200*	100
3	5	(0.1, 0.1, 0.6, 0.1, 0.1)	$\begin{bmatrix} 4 & 4 & 4 & 4 & 4 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix}$	$0.3I_2$	200*	100
4	5	(0.1, 0.1, 0.6, 0.1, 0.1)	$\begin{bmatrix} -4 & -4 & -4 & -4 & -4 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix}$	$0.3I_2$	200*	100
5	5	(0.1, 0.1, 0.6, 0.1, 0.1)	$\begin{bmatrix} -2 & -1 & 0 & 1 & 2 \\ -3 & -3 & -3 & -3 & -3 \end{bmatrix}$	$0.3I_2$	200*	100

* 20% labeling rate. I_2 represents a two-dimensional identity matrix

Algorithm 1 E^{2}ALMDA

Input: X_{L}: labeled dataset; X_{U}: unlabeled dataset; J: number of classes; t: threshold; G: number of sub-classifiers

Output: MDA models $M=\{M^g\}$ $(g=1,2,\dots,G)$; KNN classification model K

The G different training sub-datasets X_{L}^g $(g=1,2,\dots,G)$ are obtained using bagging in X_{L}
for $g=1$ to G **do**
  Initialize X_{U}
  Set the number of cycles $c=0$
  while $c=0$ or $\frac{\sum_{i=1}^{n_{\mathrm{U}}^{(\theta+1)}} \mathrm{Var}^{(\theta+1)}(p_i|g)}{n_{\mathrm{U}}^{(\theta+1)}} > \frac{\sum_{i=1}^{n_{\mathrm{U}}^{(\theta)}} \mathrm{Var}^{(\theta)}(p_i|g)}{n_{\mathrm{U}}^{(\theta)}}$ **do**
    Obtain an MDA model M^g using X_{L}^g
    On the basis of model M^g, obtain the confusion matrix C^g on X_{L} and the probability p_{ij}^g of unlabeled data x_i belonging to the j^{th} class
    According to C^g, obtain the credibility c_j^g $(j=1,2,\dots,J)$ and the insufficiency r_j^g $(j=1,2,\dots,J)$, and then calculate the entropy, confidence, and deficiency of all unlabeled data
    Try to add a different unlabeled sample from X_{U} to X_{L}^g each time, retrain M^g, and obtain the error of other unlabeled data
    Calculate the value of all unlabeled data using the integration of entropy, confidence, deficiency, and error
    Label the unlabeled sample x_{s}^g with the maximum value by M^g
    Add x_{s}^g into X_{L}^g, and remove it from X_{U}
    Update $c=c+1$
  end while
  return M^g
end for
return M
Input all the labeled data into M to obtain a set of posterior probability matrices PL
Obtain the KNN classification model K by PL
return M, K

information for unlabeled data is only given by the labeled data, which obviously represent a small part of all available data. To solve the above dilemma, active learning is exploited to add highly informative data to the training process, which can be more targeted by making class boundaries clearer. Unfortunately, only entropy and prediction error are involved in evaluating the value of each unlabeled sample. To improve the reliability and robustness of classification results, the enhanced active learning in E^{2}ALMDA introduces two new indexes, which can reduce the possibility of introducing noise into the model. Therefore, E^{2}ALMDA has the best performance among these five methods. In addition, ensemble learning is employed to avoid overfitting problems caused by one single model based on all labeled samples.

Table 3 Classification accuracy of the numerical example (80% unlabeled samples)

Class	Accuracy				
	MDA	SMDA	ALMDA	ALSemiFDA	E^{2}ALMDA
1	0.67	0.81	0.27	0.53	**0.83**
2	0.92	0.92	0.85	**0.98**	0.92
3	**1.00**	**1.00**	0.86	0.87	**1.00**
4	0.05	0.12	0.08	0.29	**0.48**
5	0.48	0.45	0.95	**0.96**	0.88
Average	0.62	0.66	0.60	0.73	**0.82**

The best results are in bold

Furthermore, the overall variance changes of ALMDA and E^{2}ALMDA are shown in Fig. 2, where each subfigure represents the variance changes using different sub-classifiers. Note that the introduction of high-value samples may lead to the loss of variance of the unlabeled dataset in the initial stage and that a small decrease in variance at the very beginning can be tolerated. According to the previous research (Yin et al., 2018, 2019), the ending threshold of

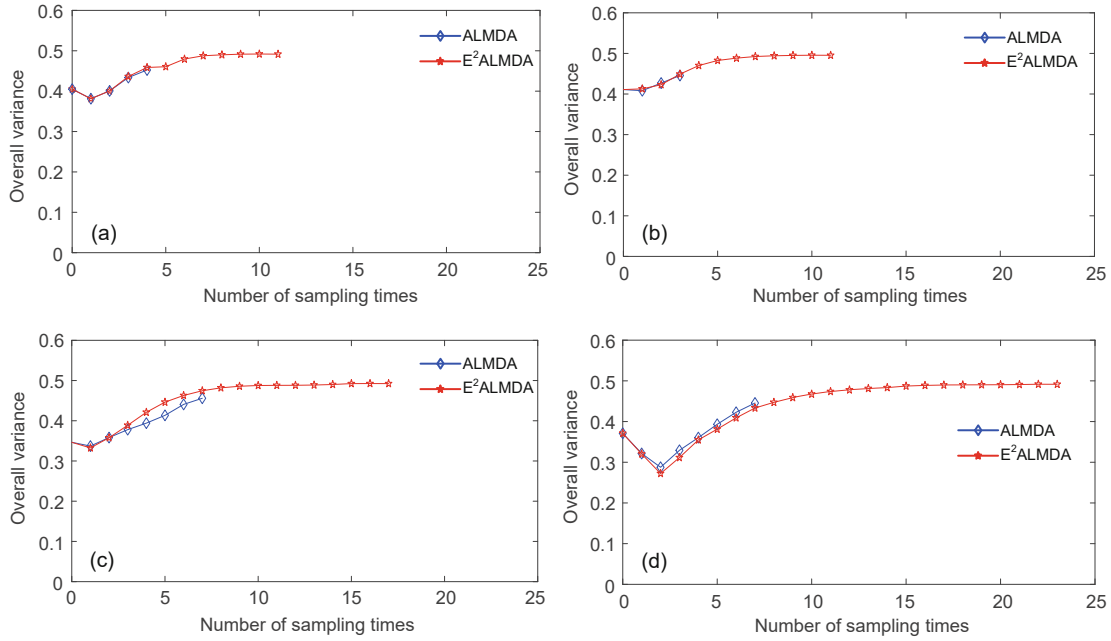

Fig. 2 Overall variance changes of ALMDA and E²ALMDA in the numerical example: (a) sub-classifier 1; (b) sub-classifier 2; (c) sub-classifier 3; (d) sub-classifier 4. Each time 5% unlabeled samples are sampled from the unlabeled dataset

active learning is manually set to 0.03. However, the variance still can be improved according to the results using the proposed E²ALMDA. The iteration of E²ALMDA stops once the maximum variance is reached, which sounds more reasonable than the predetermined threshold set in ALMDA.

4.2 Tennessee Eastman process (TEP)

TEP is a simulation model of an actual industrial process (Downs and Vogel, 1993) and has been widely used to evaluate the classification effectiveness of algorithms (Liu J et al., 2020; Zheng et al., 2020). The TEP flowchart is shown in Fig. S1 in the supplementary materials; it is composed mainly of a reactor, condenser, separator, compressor, and stripper. In the simulation process, one normal operation condition and 21 types of faults are collected, each of which contains 960 samples and 53 variables.

In this study, all variables except the agitator speed are selected for fault classification, and are listed in Tables S1 and S2 (see supplementary materials). Four sub-classifiers are initialized by the bagging technique for ensemble learning, where all 22 different types of data are used (Table S3 in the supplementary materials). Each fault appears after the 160$^{\text{th}}$ sample, where the 200$^{\text{th}}$ to 500$^{\text{th}}$ samples are chosen as training samples and the 501$^{\text{st}}$ to 600$^{\text{th}}$ samples are selected as testing samples. To simulate the semi-supervised situation, the labels of the 260$^{\text{th}}$ to 500$^{\text{th}}$ samples are removed and these samples are used as unlabeled samples, and other training samples are considered as labeled samples.

Table 4 shows that the use of unlabeled samples does have a significant impact on the improvement of classification performance. The classification results of MDA are degraded because of the scarcity of labeled samples and the absence of valuable information in unlabeled samples. Compared with the MDA method, the classification performance of SMDA is improved on some classes due to the introduction of unlabeled samples. However, all unlabeled samples are directly employed without any selection process, resulting in severe disturbance by unrelated unlabeled information. Hence, this method still has difficulty in identifying most faults. Furthermore, the labels of unlabeled data are determined mainly by the labeled data, which greatly limits the involvement of supplementary information within unlabeled data. Alternatively, entropy is adopted as a selection index in ALMDA and ALSemiFDA to introduce high-value samples from unlabeled dataset to tackle the boundary problem among classes. E²ALMDA shows better

Table 4 Accuracy of Tennessee Eastman process (TEP) classification (80% unlabeled samples)

Class	Accuracy				
	MDA	SMDA	ALMDA	ALSemiFDA	E^{2}ALMDA
Normal	0.1475	0.165	0.1125	0.1625	**0.1725**
Fault 1	**1**	**1**	**1**	**1**	**1**
Fault 2	**1**	**1**	**1**	**1**	**1**
Fault 3	0.2025	0.1875	0.1775	0.0925	**0.2375**
Fault 4	0.1325	0.1975	0.1675	0.0675	**0.3025**
Fault 5	0.2375	0.1500	0.0725	0.1350	**0.2425**
Fault 6	0.8225	0.9875	0.9825	0.9175	**1**
Fault 7	0.5175	0.3425	0.2575	0.1325	**0.6675**
Fault 8	0.7375	0.5825	0.4100	0.2500	**0.9125**
Fault 9	0.1325	0.1075	0.1125	0.0925	**0.1375**
Fault 10	0.1525	0.1425	0.0925	0.1125	**0.1875**
Fault 11	0.3200	0.3200	0.2125	0.0650	**0.3400**
Fault 12	0.6875	0.5225	0.3500	0.2300	**0.7225**
Fault 13	0.7575	0.7125	0.5875	0.5225	**0.9625**
Fault 14	0.7725	0.7825	0.7825	0.3550	**0.8925**
Fault 15	0.1125	0.1875	**0.2025**	0.0325	0.1875
Fault 16	0.1425	0.1100	0.1075	0.1450	**0.1475**
Fault 17	0.7225	0.6225	0.5225	0.4525	**0.7375**
Fault 18	0.7825	0.7675	0.7500	0.5025	**0.7875**
Fault 19	0.3325	**0.4525**	0.4275	0.0925	0.3775
Fault 20	0.4125	0.3925	0.2725	0.0375	**0.4675**
Fault 21	0.1725	0.2825	0.1625	0.1225	**0.3125**
Average	0.4681	0.4552	0.3983	0.2964	**0.5361**

The best results are in bold

classification performance than the other four methods because the classification performance of each sub-classifier is effectively improved owing to the introduction of enhanced active learning. The newly designed indexes can significantly reduce the consequences caused by noise, while the value judgment of unlabeled samples becomes more reasonable. Finally, ensemble learning improves the robustness and ability to tolerate noise of the semi-supervised model through enhanced weak classifiers and corresponding ensemble rules.

Similarly, Fig. 3 shows the overall variance changes of ALMDA and E^{2}ALMDA under four different sub-classifiers, in which the prediction error threshold of ALMDA is set to 0.18. Compared with ALMDA, E^{2}ALMDA can collect more valuable information from the unlabeled dataset until the overall maximum variance is reached.

5 Conclusions

In this paper, the E^{2}ALMDA method has been proposed to provide a comprehensive explanation for the semi-supervised fault classification issue. First, to improve the robustness of the proposed method, several sub-datasets have been collected from the original dataset through the bagging technique, where corresponding weak classifiers were established. Second, several new indexes have been designed to introduce suitable unlabeled samples in the evolution of sub-classifiers. Instead of human labeling, we have proposed a model labeling solution for unlabeled samples in which a final criterion and an unlabeled sample value evaluation index were carried out based on the newly designed indexes to guarantee the performance of each weak classifier. Third, the results of enhanced sub-classifiers have been integrated using the KNN method to obtain the final fault classification results. Finally, the performance of the proposed method has been further evaluated using a numerical example and the TEP benchmark in which the superiority of the proposed method was shown.

Contributors

Weijun WANG designed the research. Weijun WANG and Yuchen HE processed the data. Weijun WANG drafted the paper. Yuchen HE and Xinyun FANG helped organize

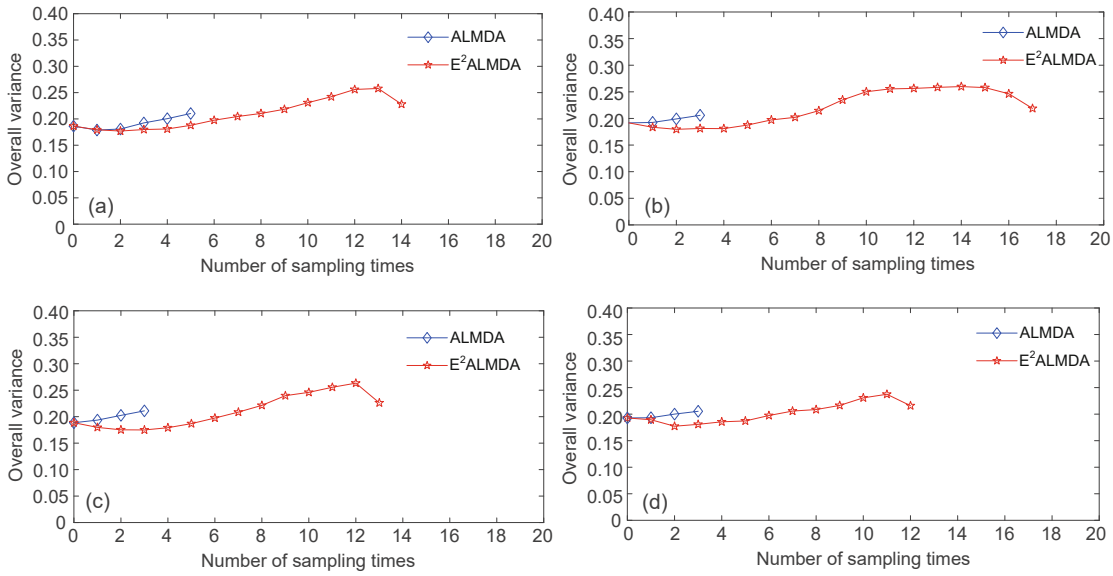

Fig. 3 Overall variance changes of ALMDA and E²ALMDA in the Tennessee Eastman process: (a) sub-classifier 1; (b) sub-classifier 2; (c) sub-classifier 3; (d) sub-classifier 4. Each time 5% unlabeled samples are sampled from the unlabeled dataset

the paper. Yun WANG and Jun WANG revised and finalized the paper.

Acknowledgements

The authors would like to thank Huanghui ZHANG from Fujian Institute of Metrology for helping improve this paper.

Compliance with ethics guidelines

Weijun WANG, Yun WANG, Jun WANG, Xinyun FANG, and Yuchen HE declare that they have no conflict of interest.

References

Abellán J, Masegosa AR, 2010. Bagging decision trees on data sets with classification noise. Proc 6th Int Symp on Foundations of Information and Knowledge Systems, p.248-265. https://doi.org/10.1007/978-3-642-11829-6_17

Abramson N, Braverman D, Sebestyen G, 1963. Pattern recognition and machine learning. *IEEE Trans Inform Theory*, 9(4):257-261. https://doi.org/10.1109/TIT.1963.1057854

Araya DB, Grolinger K, ElYamany HF, et al., 2017. An ensemble learning framework for anomaly detection in building energy consumption. *Energy Build*, 144:191-206. https://doi.org/10.1016/j.enbuild.2017.02.058

Blum A, Chawla S, 2001. Learning from labeled and unlabeled data using graph mincuts. Proc 18th Int Conf on Machine Learning, p.19-26.

Botre C, Mansouri M, Karim MN, et al., 2017. Multiscale PLS-based GLRT for fault detection of chemical processes. *J Loss Prev Process Ind*, 46:143-153. https://doi.org/10.1016/j.jlp.2017.01.008

Bouveyron C, Girard S, 2009. Robust supervised classification with mixture models: learning from data with uncertain labels. *Patt Recogn*, 42(11):2649-2658. https://doi.org/10.1016/j.patcog.2009.03.027

Chapelle O, Sindhwani V, Sathiya Keerthi S, 2006. Branch and bound for semi-supervised support vector machines. Proc 19th Int Conf on Neural Information Processing Systems, p.217-224. https://doi.org/10.5555/2976456.2976484

Chen X, Wang ZP, Zhang Z, et al., 2018. A semi-supervised approach to bearing fault diagnosis under variable conditions towards imbalanced unlabeled data. *Sensors*, 18(7):2097. https://doi.org/10.3390/s18072097

Chiang LH, Russell EL, Braatz RD, 2000. Fault diagnosis in chemical processes using Fisher discriminant analysis, discriminant partial least squares, and principal component analysis. *Chemom Intell Lab Syst*, 50(2):243-252. https://doi.org/10.1016/S0169-7439(99)00061-1

Chiang LH, Kotanchek ME, Kordon AK, 2004. Fault diagnosis based on Fisher discriminant analysis and support vector machines. *Comput Chem Eng*, 28(8):1389-1401. https://doi.org/10.1016/j.compchemeng.2003.10.002

Cui XD, Huang J, Chien JT, 2012. Multi-view and multi-objective semi-supervised learning for HMM-based automatic speech recognition. *IEEE Trans Audio Speech Lang Process*, 20(7):1923-1935. https://doi.org/10.1109/TASL.2012.2191955

Deng XG, Liu XY, Cao YP, et al., 2022. Incipient fault detection for dynamic chemical processes based on enhanced CVDA integrated with probability information and fault-sensitive features. *J Process Contr*, 114:29-41. https://doi.org/10.1016/j.jprocont.2022.04.001

Dietterich TG, 2000. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Mach Learn*, 40(2):139-157. https://doi.org/10.1023/A:1007607513941

Dong YN, Qin SJ, 2018. A novel dynamic PCA algorithm for dynamic data modeling and process monitoring. *J Process Contr*, 67:1-11. https://doi.org/10.1016/j.jprocont.2017.05.002

Downs JJ, Vogel EF, 1993. A plant-wide industrial process control problem. *Comput Chem Eng*, 17(3):245-255. https://doi.org/10.1016/0098-1354(93)80018-I

Farajzadeh-Zanjani M, Hallaji E, Razavi-Far R, et al., 2021. Adversarial semi-supervised learning for diagnosing faults and attacks in power grids. *IEEE Trans Smart Grid*, 12(4):3468-3478. https://doi.org/10.1109/TSG.2021.3061395

Feng J, Wang J, Han ZY, 2013. Process monitoring for chemical process based on semi-supervised principal component analysis. Proc 25th Chinese Control and Decision Conf, p.4282-4286. https://doi.org/10.1109/CCDC.2013.6561704

Fraley C, Raftery AE, 2002. Model-based clustering, discriminant analysis, and density estimation. *J Am Stat Assoc*, 97(458):611-631. https://doi.org/10.1198/016214502760047131

Ge ZQ, 2016. Supervised latent factor analysis for process data regression modeling and soft sensor application. *IEEE Trans Contr Syst Technol*, 24(3):1004-1011. https://doi.org/10.1109/TCST.2015.2473817

Ge ZQ, 2017. Review on data-driven modeling and monitoring for plant-wide industrial processes. *Chemom Intell Lab Syst*, 171:16-25. https://doi.org/10.1016/j.chemolab.2017.09.021

Ge ZQ, 2018. Process data analytics via probabilistic latent variable models: a tutorial review. *Ind Eng Chem Res*, 57(38):12646-12661. https://doi.org/10.1021/acs.iecr.8b02913

Ge ZQ, Song ZH, Gao FR, 2013. Review of recent research on data-based process monitoring. *Ind Eng Chem Res*, 52(10):3543-3562. https://doi.org/10.1021/ie302069q

Ge ZQ, Song ZH, Ding SX, et al., 2017. Data mining and analytics in the process industry: the role of machine learning. *IEEE Access*, 5:20590-20616. https://doi.org/10.1109/ACCESS.2017.2756872

Hady MFA, Schwenker F, 2010. Combining committee-based semi-supervised learning and active learning. *J Comput Sci Technol*, 25(4):681-698. https://doi.org/10.1007/s11390-010-9357-6

Harkat MF, Mansouri M, Nounou MN, et al., 2019. Fault detection of uncertain chemical processes using interval partial least squares-based generalized likelihood ratio test. *Inform Sci*, 490:265-284. https://doi.org/10.1016/j.ins.2019.03.068

Hastie T, Tibshirani R, 1996. Discriminant analysis by Gaussian mixtures. *J Roy Stat Soc Ser B*, 58(1):155-176. https://doi.org/10.1111/j.2517-6161.1996.tb02073.x

He YL, Li K, Zhang N, et al., 2021. Fault diagnosis using improved discrimination locality preserving projections integrated with sparse autoencoder. *IEEE Trans Instrum Meas*, 70:3527108. https://doi.org/10.1109/TIM.2021.3125975

Huang CC, Chen T, Yao Y, 2013. Mixture discriminant monitoring: a hybrid method for statistical process monitoring and fault diagnosis/isolation. *Ind Eng Chem Res*, 52(31):10720-10731. https://doi.org/10.1021/ie400418c

Ipeirotis PG, Provost F, Wang J, 2010. Quality management on Amazon Mechanical Turk. Proc ACM SIGKDD Workshop on Human Computation, p.64-67. https://doi.org/10.1145/1837885.1837906

Jin YR, Qin CJ, Huang YX, et al., 2021. Actual bearing compound fault diagnosis based on active learning and decoupling attentional residual network. *Measurement*, 173:108500. https://doi.org/10.1016/j.measurement.2020.108500

Kalantar B, Al-Najjar HAH, Pradhan B, et al., 2019. Optimized conditioning factors using machine learning techniques for groundwater potential mapping. *Water*, 11(9):1909. https://doi.org/10.3390/w11091909

Liu J, Song CY, Zhao J, 2018. Active learning based semi-supervised exponential discriminant analysis and its application for fault classification in industrial processes. *Chemom Intell Lab Syst*, 180:42-53. https://doi.org/10.1016/j.chemolab.2018.07.003

Liu J, Song CY, Zhao J, et al., 2020. Manifold-preserving sparse graph-based ensemble FDA for industrial label-noise fault classification. *IEEE Trans Instrum Meas*, 69(6):2621-2634. https://doi.org/10.1109/TIM.2019.2930157

Liu JW, Liu Y, Luo XL, 2015. Semi-supervised learning methods. *Chin J Comput*, 38(8):1592-1617 (in Chinese). https://doi.org/10.11897/SP.J.1016.2015.01592

Liu Y, Ge ZQ, 2018. Weighted random forests for fault classification in industrial processes with hierarchical clustering model selection. *J Process Contr*, 64:62-70. https://doi.org/10.1016/j.jprocont.2018.02.005

MacGregor J, Cinar A, 2012. Monitoring, fault diagnosis, fault-tolerant control and optimization: data driven methods. *Comput Chem Eng*, 47:111-120. https://doi.org/10.1016/j.compchemeng.2012.06.017

Pu XK, Li CG, 2021. Probabilistic information-theoretic discriminant analysis for industrial label-noise fault diagnosis. *IEEE Trans Ind Inform*, 17(4):2664-2674. https://doi.org/10.1109/TII.2020.3001335

Raina R, Battle A, Lee H, et al., 2007. Self-taught learning: transfer learning from unlabeled data. Proc 24th Int Conf on Machine Learning, p.759-766. https://doi.org/10.1145/1273496.1273592

Raykar VC, Yu SP, Zhao LH, et al., 2010. Learning from crowds. *J Mach Learn Res*, 11:1297-1322. https://doi.org/10.5555/1756006.1859894

Schwenker F, Trentin E, 2014. Pattern classification and clustering: a review of partially supervised learning approaches. *Patt Recogn Lett*, 37:4-14. https://doi.org/10.1016/j.patrec.2013.10.017

Settles B, 2012. Active Learning. Morgan & Claypool Publishers, USA. https://doi.org/10.2200/S00429ED1V01Y201207AIM018

Shao WM, Tian XM, 2017. Semi-supervised selective ensemble learning based on distance to model for nonlinear soft sensor development. *Neurocomputing*, 222:91-104. https://doi.org/10.1016/j.neucom.2016.10.005

Shao WM, Ge ZQ, Song ZH, 2019a. Semi-supervised mixture of latent factor analysis models with application to online key variable estimation. *Contr Eng Pract*, 84:32-47. https://doi.org/10.1016/j.conengprac.2018.11.008

Shao WM, Ge ZQ, Song ZH, et al., 2019b. Nonlinear industrial soft sensor development based on semi-supervised probabilistic mixture of extreme learning machines. *Contr Eng Pract*, 91:104098. https://doi.org/10.1016/j.conengprac.2019.07.016

Snow R, O'Connor B, Jurafsky D, et al., 2008. Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. Proc Conf on Empirical Methods in Natural Language Processing, p.254-263.

Wang J, Feng J, Han ZY, 2014. Fault detection for the class imbalance problem in semiconductor manufacturing processes. *J Circ Syst Comput*, 23(4):1450049. https://doi.org/10.1142/S0218126614500492

Wang JB, Shao WM, Song ZH, 2019. Semi-supervised variational Bayesian student's t mixture regression and robust inferential sensor application. *Contr Eng Pract*, 92:104155. https://doi.org/10.1016/j.conengprac.2019.104155

Wang L, Tian H, Zhang H, 2021. Soft fault diagnosis of analog circuits based on semi-supervised support vector machine. *Analog Integr Circ Signal Process*, 108(2):305-315. https://doi.org/10.1007/s10470-021-01851-w

Yan ZB, Huang CC, Yao Y, 2014. Semi-supervised mixture discriminant monitoring for chemical batch processes. *Chemom Intell Lab Syst*, 134:10-22. https://doi.org/10.1016/j.chemolab.2014.03.002

Yao L, Ge ZQ, 2017. Locally weighted prediction methods for latent factor analysis with supervised and semisupervised process data. *IEEE Trans Autom Sci Eng*, 14(1):126-138. https://doi.org/10.1109/TASE.2016.2608914

Yin LL, Wang HG, Fan WH, et al., 2018. Combining active learning and Fisher discriminant analysis for the semi-supervised process monitoring. *IFAC-PapersOnLine*, 51(21):147-151. https://doi.org/10.1016/j.ifacol.2018.09.407

Yin LL, Wang HG, Fan WH, et al., 2019. Incorporate active learning to semi-supervised industrial fault classification. *J Process Contr*, 78:88-97. https://doi.org/10.1016/j.jprocont.2019.04.008

Yuen MC, King I, Leung KS, 2011. A survey of crowdsourcing systems. Proc IEEE 3rd Int Conf on Privacy, Security, Risk and Trust and IEEE 3rd Int Conf on Social Computing, p.766-773. https://doi.org/10.1109/PASSAT/SocialCom.2011.203

Zaman SMK, Liang XD, 2021. An effective induction motor fault diagnosis approach using graph-based semi-supervised learning. *IEEE Access*, 9:7471-7482. https://doi.org/10.1109/ACCESS.2021.3049193

Zhang N, Xu Y, Zhu QX, et al., 2022. Improved locality preserving projections based on heat-kernel and cosine weights for fault classification in complex industrial processes. *IEEE Trans Reliab*, early access. https://doi.org/10.1109/TR.2021.3139539

Zheng JH, Wang HJ, Song ZH, et al., 2019. Ensemble semi-supervised Fisher discriminant analysis model for fault classification in industrial processes. *ISA Trans*, 92:109-117. https://doi.org/10.1016/j.isatra.2019.02.021

Zheng JH, Zhu JL, Chen GJ, et al., 2020. Dynamic Bayesian network for robust latent variable modeling and fault classification. *Eng Appl Artif Intell*, 89:103475. https://doi.org/10.1016/j.engappai.2020.103475

Zhong K, Han M, Qiu T, et al., 2020. Fault diagnosis of complex processes using sparse kernel local Fisher discriminant analysis. *IEEE Trans Neur Netw Learn Syst*, 31(5):1581-1591. https://doi.org/10.1109/TNNLS.2019.2920903

Zou Y, Yu ZD, Liu XF, et al., 2019. Confidence regularized self-training. Proc IEEE/CVF Int Conf on Computer Vision, p.5981-5990. https://doi.org/10.1109/ICCV.2019.00608

List of supplementary materials

Fig. S1 Tennessee Eastman process (TEP)

Table S1 Selected measured variables for monitoring

Table S2 Selected manipulated variables for monitoring

Table S3 Twenty-two operating conditions simulated in TEP