Weijun WANG, Yun WANG, Jun WANG, Xinyun FANG, Yuchen HE, 2022. Ensemble enhanced active learning mixture discriminant analysis model and its application for semi-supervised fault classification. *Frontiers of Information Technology & Electronic Engineering,* 23(12):1814-1827. https://doi.org/10.1631/FITEE.2200053

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

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

Corresponding author: Yuchen HE
E-mail: yche@cjlu.edu.cn
ORCID: https://orcid.org/0000-0002-0528-2778

Motivation

1. Fault classification plays a significant role in guaranteeing the effective operation of process monitoring systems. Meanwhile, with the rapid development of information technology, massive valuable data have been collected, transmitted, and stored. Consequently, data-driven methods have recently become feasible methods for fault classification.
2. 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.
3. 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.

Main idea

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, which improves the reliability of model labeling through the four proposed indexes, while avoiding human interference.
3. A reasonable stopping criterion is proposed to introduce more informative samples to the sub-dataset.

Method

1. The bagging technique is introduced to randomly collect labeled samples to construct sub-classifiers for subsequent integration. Due to a sophisticated process, the data collections are unlikely to follow Gaussian distributions. Therefore, the mixture discriminant analysis (MDA) technique is adopted where the data distribution is regarded as a mixture of multiple Gaussian components.

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

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

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

Method (Cont'd)

2. Two new selection indexes, called confidence and deficiency, are designed in this study to realize model labeling for unlabeled samples. Meanwhile, the inherent supplementary information within the unlabeled dataset can be extracted by traditional active learning methods.

$$\text{Confidence}(i|g) = \sum_{j=1}^{J} \left(p_{ij}^{g} c_{j}^{g}\right)$$

$$\text{Deficiency}(i|g) = \sum_{j=1}^{J} \left(p_{ij}^{g} r_{j}^{g}\right)$$

$$e_{i}^{g} = -\sum_{j=1}^{J} \left(p_{ij}^{g} \ln p_{ij}^{g}\right)$$

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

Method (Cont'd)

3. After each sub-classifier is strengthened, ensemble learning will be used to strengthen the robustness of the classification results by integrating the performance of all sub-classifiers.

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

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

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

Method (Cont'd)

4. The detailed E^{2}ALMDA flowchart is shown in Fig. 1.

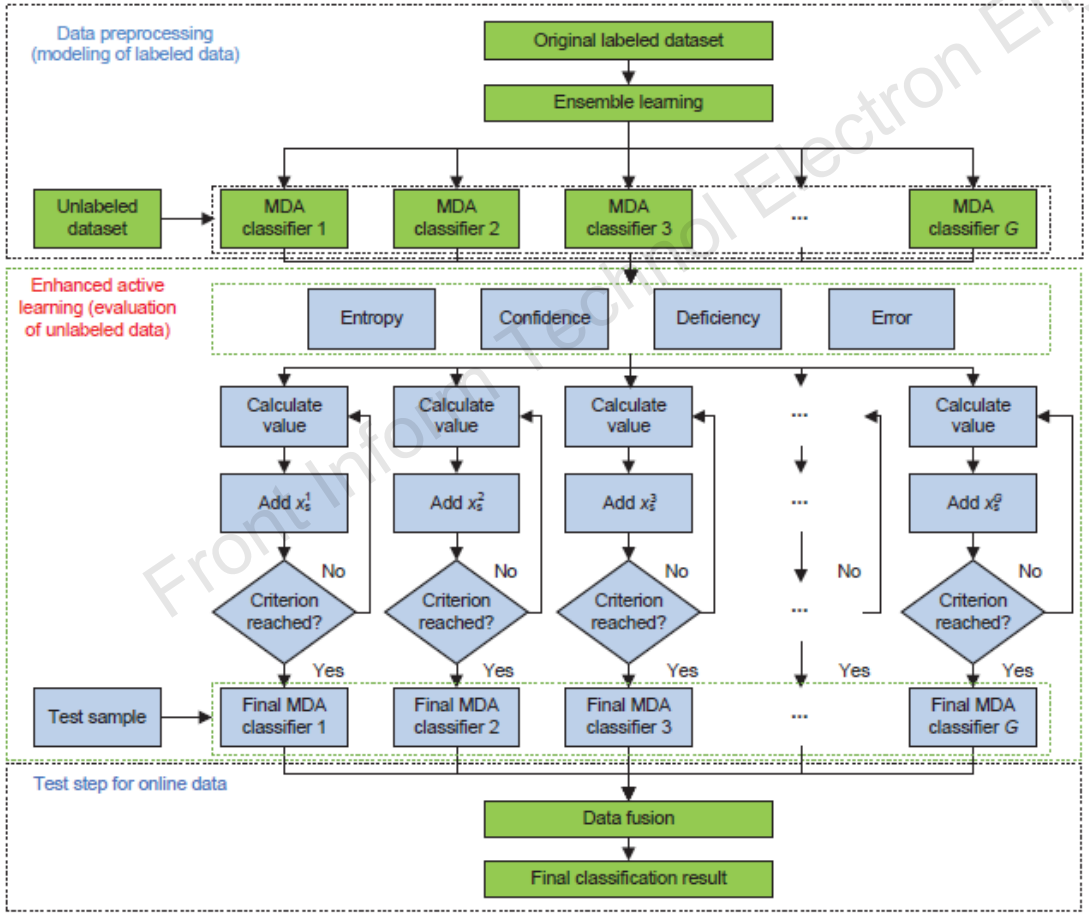


Fig. 1 Flowchart of E^{2}ALMDA modeling and testing

Major results

Classification results

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

Class	Accuracy				
	MDA	SMDA	ALMDA	ALSemiFDA	E^2ALMDA
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

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

Class	Accuracy				
	MDA	SMDA	ALMDA	ALSemiFDA	E^2ALMDA
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

Major results (Cont'd)

Overall variance changes

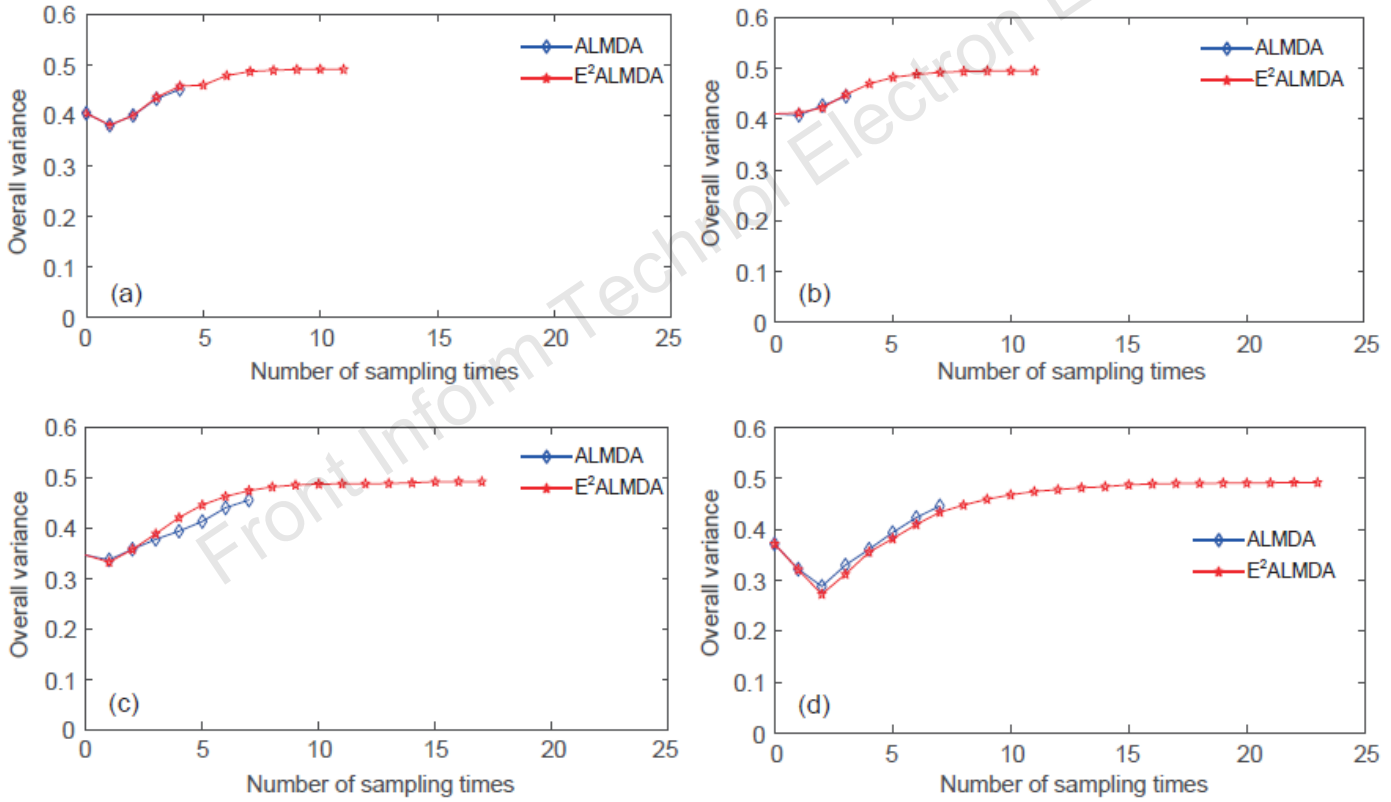


Fig. 2 Overall variance changes of ALMDA and E²ALMDA in the numerical example: (a) sub-classifier 1; (b) sub-classifier 2; (c) sub-classifier 3; (d) sub-classifier 4. Each time 5% unlabeled samples are sampled from the unlabeled dataset

Major results (Cont'd)

Overall variance changes

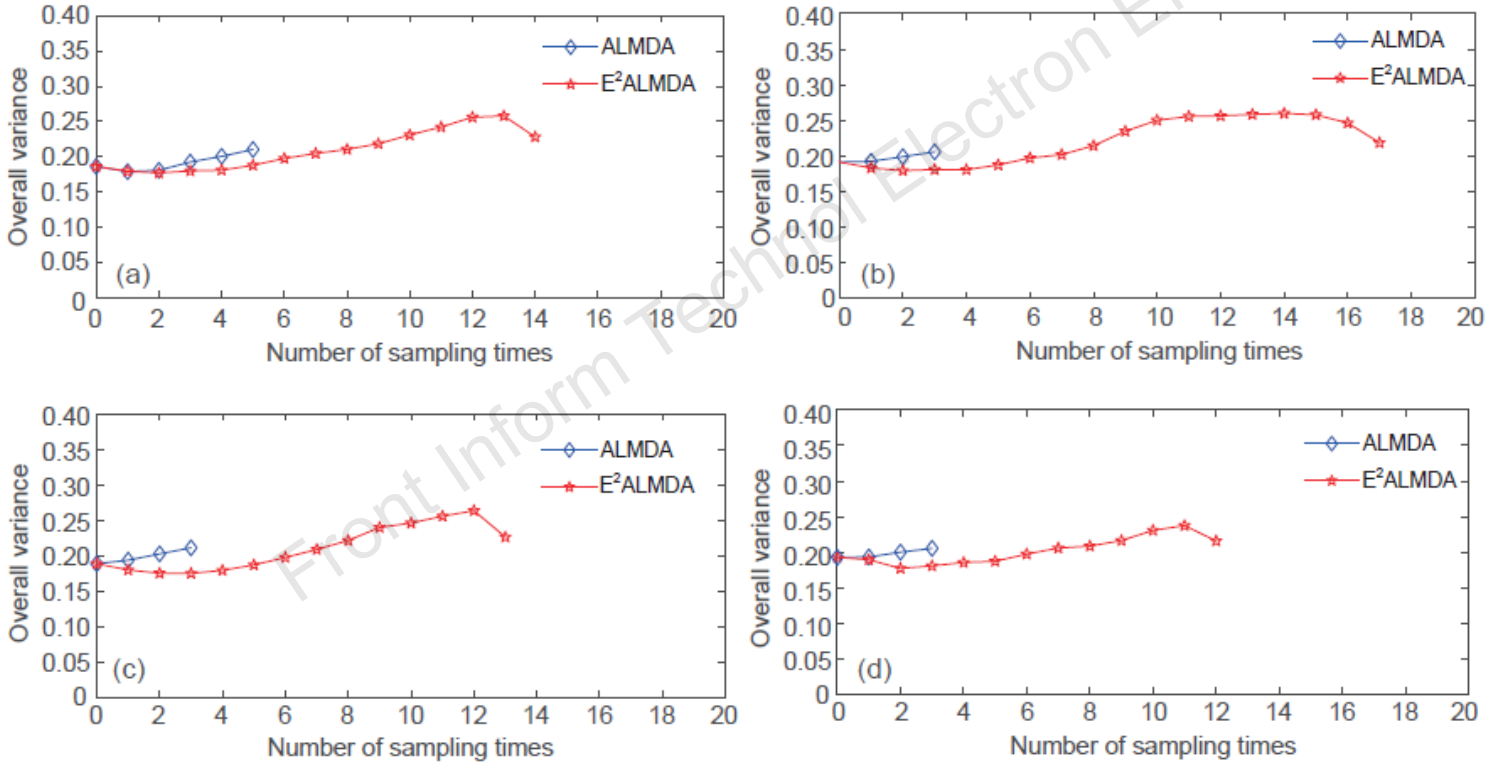


Fig. 3 Overall variance changes of ALMDA and E²ALMDA in the Tennessee Eastman process: (a) sub-classifier 1; (b) sub-classifier 2; (c) sub-classifier 3; (d) sub-classifier 4. Each time 5% unlabeled samples are sampled from the unlabeled dataset

Conclusions

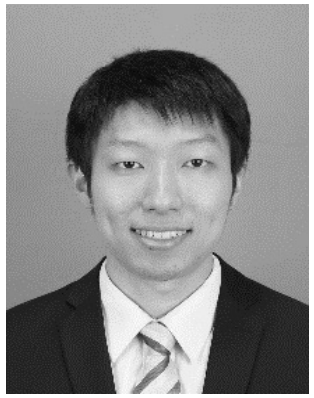
1. To improve the robustness of the proposed method, several sub-datasets have been collected from the original dataset through the bagging technique, where corresponding weak classifiers were established.
2. Several new indexes have been designed to introduce suitable unlabeled samples in the evolution 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.
3. The results of enhanced sub-classifiers have been integrated using the *K*-nearest neighbor (KNN) method to obtain the final fault classification results.

Weijun WANG is currently pursuing the MS degree with the College of Mechanical and Electrical Engineering, China Jiliang University. His research interests include mixture model, fault detection, fault classification and their applications in industrial processes.

Yun WANG received the BS degree from Zhejiang University of Technology, in 2010 and the MS degree from Zhejiang University of Technology, in 2012. She is currently a Faculty Member of Zhejiang Tongji Vocational College of Science and Technology, where she is currently a Lecturer with the Mechanical and Electrical Engineering Department. Her research interests include fault diagnosis, data analysis, and their applications in industrial processes.

Jun WANG received the BS and MS degrees from Wuhan University respectively in 1991 and 2001, and the PhD degree from Huazhong University of Science and Technology, China, in 2013. He is a Faculty Member of China Jiliang University, where he is currently a Lecturer with the College of Mechanical and Electrical Engineering. His research interests include fuel flow and combustion, as well as solar power system.

Xinyun FANG received the MS degree from Southeast University, China, in 2010. He is currently a member in Suzhou Institute of Metrology, where he is currently the minister of Science and Technology Information Department. His research interests include science and technology management, metering digitization, and temperature measurement.



Yuchen HE received the BS degree from Zhejiang University of Technology, in 2010 and the PhD degree from Zhejiang University, China, in 2017. He is currently a Faculty Member of China Jiliang University, where he is currently an associate professor with the College of Mechanical and Electrical Engineering. His research interests include multivariate statistical process monitoring, fault diagnosis, data analysis, and their applications in industrial processes.