Data-driven state monitoring of air preheater using density peaks clustering and evidential K-nearest neighbour classifier

Peng Sha[1], *Xiao* Wu[1], *Jiong* Shen[1,*], *Xichui* Liu[1], and *Meihong* Wang[2]

[1]Key Laboratory of Energy Thermal Conversion and Control, Ministry of Education, School of Energy and Environment, Southeast University, NanJing, 210096, P.R. China
[2]Department of Chemical and Biological Engineering, University of Sheffield, Sheffield S1 3JD, UK

Abstract. Data-driven state monitoring requiring a little priori knowledge plays a key role for timely fault detection and is therefore of great importance for the safe and economical operation of the thermal power plant. The main drawback for most of the existing data-driven methods is the complex procedure of data preprocessing and model training especially when unlabelled operating data is used. To overcome this issue, this paper proposes a new framework of data-driven state monitoring approach for the thermal power plant devices. The approach is composed of two steps. In the first step, density peaks clustering(DPC) is performed on the historical data to generate labels for the data. Then in the second step, evidential K-nearest neighbour(EKNN) method is used to monitor the current state based on the labelled historical data and operating data. Verifications on operating data of an air preheater system of a 1000MW thermal power plant show that the proposed method can identify various air leakage states accurately and efficiently.

1 Introduction

Many devices in power plants such as the pumps, fans and heaters are subject to breakdowns and malfunctions during their operating processes, which can have great impacts on their performance or even lead to considerable economic losses[1]. The conventional prevention methods, such as the scheduled or breakdown maintenance cannot provide a timely and efficient trouble detection. Consequently, how to attain state monitoring while the devices are still operating has become important and essential issues for both industrial practitioners and academic researchers. In [2]-[5], various data-driven state monitoring approaches, such as, support vector machine[2], artificial neural network[3], neuro-fuzzy systems[4] and multivariate statistical process control[5], etc. have been well developed for condition monitoring and fault detection which can avoid the complex mechanism modeling procedure, However, some problems exist and limit the application of these approach:

* Corresponding author: shenj@seu.edu.cn

(1) The training data need to be well-prepared, which greatly relies on human knowledge and is time consuming; (2) The system noise, widespread in the industrial processes, can have significant influence on the prediction accuracy of the models.

To overcome the issues mentioned above, this paper employs Density Peaks Clustering(DPC)[6] and evidential k-nearest neighbour classifier(EKNN)[7] techniques to achieve state monitoring for the air preheater of thermal power plant. Owing to the features that can find the number of clusters and detect clusters of non-convex intuitively from the data set, DPC provides an effective way in operating data pre-processing, which can find existing states of devices without making assumptions on the number of possible states[8]. EKNN provides a nonparametric procedure for the assignment of a class label to the input pattern based on the class labels of its kth nearest neighbours[9]. The KNN-based classifier can alleviate the influence of local error data points. Moreover the evidence theory can further improve the classification performance in the presence of imprecise and ambiguous information[10].

The remainder of this paper is organized as follows, the details of the proposed state monitoring method are presented in section 2. Section 3 briefly recalls the leakage of rotary air preheater. In section 4, a example of state monitoring on air preheater is presented to illustrate the effectiveness of proposed method. The last section concludes the paper.

2 Methodology

In this section, the details of the proposed state monitoring method are briefly introduced for the consequent work.

2.1 The proposed state monitoring method

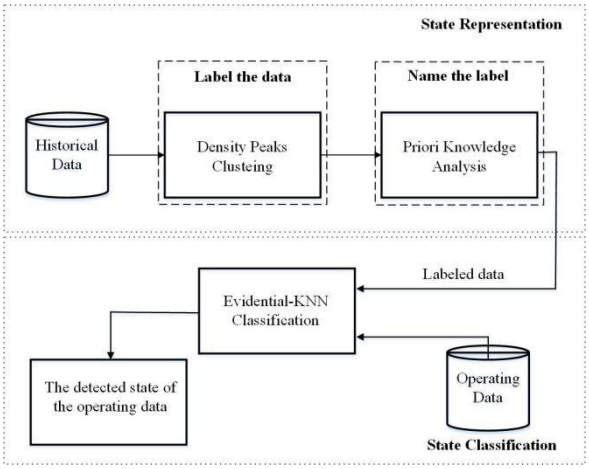

Fig. 1. State monitoring flow chart.

As shown in Figure 1, the method is composed of two steps. In the first step, DPC is performed on the historical data and the clustering result is adopted as the labels of the data. DPC can intuitively find the number of clusters and detect clusters of non-convex, which can thus find existing states of devices without making assumptions on the number of possible states. Once the clustering results are obtained, each label can be named according to the actual state based on a priori knowledge. The labeled data is then used as a baseline to predict the state of the operating data in the second step. To give a correct classification for the ambiguous states, E-KNN is adopted to identify the current data's categories based on the labeled historical data and operating data.

2.2 Density peaks clustering(DPC)

The approach of DPC is developed based on two intuitive assumptions: (i) Cluster centers are surrounded by neighbors with lower local density; and (ii) Cluster centers are selected from the region with higher local density and the distances between cluster centers are expected to be as large as possible. In order to find the potential cluster centers on the decision graph, the local density ρ_i (1) and the distance from points of higher density δ_i are calculated for each point-*i* (2).

$$\rho_i = \sum_j \chi\left(d_{ij} - d_c\right) \tag{1}$$

$$\delta_i = \begin{cases} \min_{j:\rho_i<\rho_j} d_{ij}, if\ \exists j\ s.t. \rho_i < \rho_j \\ \max_j (d_{ij}),\ otherwise \end{cases} \tag{2}$$

where d_{ij} is the distance between points *i* and *j*, and d_c is a cutoff distance which is the only variable set by users. $\chi(t)=1$ if $t<0$ and $\chi(t)=0$ otherwise. According to Rodriguez and Liao [6], one can select d_c in such a way that the average number of neighbors of each object is around 1-2% of the total number of objects in the dataset. After ρ_i and δ_i are calculated, two other steps need to be carried out for the DPC: First, Setting up the decision graph with the collection of points (ρ_i,,δ_i) and second selecting the potential cluster centers which has the highest δ and ρ. Then remaining points is assigned to the same cluster as its nearest neighbor of higher density in a single step.

2.3 Evidential K-nearest neighbour classifier(EKNN)

The algorithm of EKNN is developed based on the information provided by a training set $T = \left\{\left(x_i, m_i\right) \middle| i = 1, 2, ..., n\right\}$,of p-dimensional pattern *x* and their corresponding class label belief structures *m*, taking values in Ω. A discounting function ϕ is adopted to quantize the similarity between *x* and x_i :

$$\phi\left(d_i\right) = e^{-\gamma d_i^2} \tag{3}$$

where d_{ij} is the distance metric between *x* and x_i, $\gamma>0$, is a discounting factor to adjust the attenuation characteristic of ϕ.

The information provided by training vector x_i is postulated to induce a basic belief assignment $m\left(\cdot \middle| x_i\right)$ over Ω defined by

$$m\left(A \middle| x_i\right) = \begin{cases} \alpha\phi\left(d_i\right), & A = \left\{w_q\right\} \\ 1 - \alpha\phi\left(d_i\right), & A = \Omega \\ 0, & A \in 2^{\Omega} \setminus \left\{\Omega, \left\{\omega_q\right\}\right\} \end{cases} \tag{4}$$

where $\alpha \in (0,1)$, representing the credibility of the training vector; $m\left(A \middle| \Omega\right)$ is the belief assigned to the degree of ignorance.

The basic belief assignments of the k-nearest neighbours are combined by using the Dempster's rule of combination [7] to form a resulting BBA m :

$$m = \bigoplus_{i \in I_k} m\left(\cdot \middle| x_i\right) \tag{5}$$

where $I_k = \left\{i_1, i_2, ..., i_k\right\}$ are the indexes of the k-nearest neighbours of x in *T*.

3 The leakage of rotary air preheater

Rotary air preheater is used to recover waste heat from the exhaust flue gas in thermal power plants. It can be arranged in a large number of heating exchange surfaces with a small volume and can be easily maintained and operated [11]. Air preheaters have a general operating fault known as leakage[12], which can be categorized as direct leakage and entrained leakage. Direct leakage occurs when higher pressure air leaks into the lower pressure flue gas through gaps between the rotating and stationary parts. According to Bernoulli equation[13]. Direct leakage F_w can be calculated by Eq.6.

$$F_w = F_1 kA\sqrt{2(p_1 - p_2)\rho_1} \tag{6}$$

where F_1 is the modifying factor that corrects the deviation between the operation under laboratory testing and in the actual operating conditions of the considered air preheater in the power plants; k is the orifice coefficient determined by the wind tunnel test; A is the gap section area; p_1,p_2 are the pressure of higher side and lower side; ρ_1 is the density of higher side.

Entrained leakage can be described as being a result of the rotation of the matrix from one stream to the next[14]. Air is carried into the flue gas stream as the heating surface components or baskets are rotated from the air stream to the flue gas stream. This leakage is directly proportional to the void volume of the rotor and the motor speed[15].Air leakage can cause a significant drop in the effectiveness of an air preheater, for every 1% of leakage on the cold end, the drop in air preheater effectiveness was approximately 7%[13]. Therefore, it is of great significance to monitor the state of air leakage accurately in real time and adjust the maintenance timely. Typically 70%-80% of the total air leakage comes from the direct leakage. To simplify the problem, this paper only focus on the change of direct air leakage.

4 An example of real operating data

This section provides an example on leakage state monitoring of air preheater system to test the effectiveness of the proposed method. An operating data set of 5000 minutes from a 1000MWe coal-fired power plant is used to test the proposed method. According on them, two steps of data preprocessing are then carried out on the data: (1) All samples with null or uncertain values were replaced by sliding average; (2) The data were smoothed..

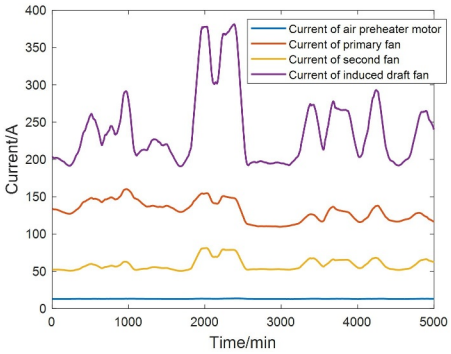

Fig. 2. Operating data for current of air preheater.

Four variables which are bound up with the operating condition is used to analyze the leakage, namely, the current of air preheater motor, primary fan, second fan and the induced draft fan. As shown in Figure 2, the current of air preheater motor is almost unchanged, therefore, we can ignore the influence of entrained leakage in this case. The

current of primary fan, second fan and induced draft fan vary with time. Among them, the current of primary fan and second fan have similar trend of change, while the current of induced draft fan is different, indicating the condition of leakage has changed during the time.

4.1 Clustering based data labelling

According to the above analysis and relevant expertise, six key operating variables are selected to monitor the operating condition of the air preheater, namely, the outlet pressure of primary fan ,the outlet pressure of second fan, the inlet flue gas pressure of air preheater, the Outlet flue gas pressure of air preheater, the outlet second air pressure of air preheater, the outlet primary air pressure of air preheater.

Cluster analysis is a kind of unsupervised learning, and we need to analyze its results to verify its rationality. As mentioned above the current of air preheater motor keeping the same in this period, and the entrained leakage is assumed to keep the same. According to Eq.(6), the leakage is proportional to the square root of the pressure difference. In this case, we give an index defined in Eq.(7) to qualitatively describe the condition of air leakage and help to interpret the clustering results.

$$\Delta P = \Delta P_1 + \Delta P_2 \tag{7}$$

where Δp_1 is the difference between the outlet pressure of primary fan and the outlet primary air pressure of air preheater, Δp_1 is the difference between the outlet pressure of second fan and the outlet second air pressure of air preheater, ΔP qualitatively reflects the total air leakage of air preheater.

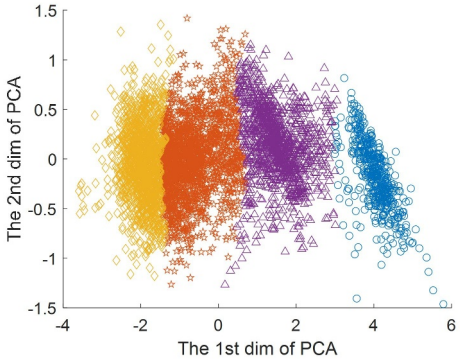

Fig. 3. Clustering result of operating data for air preheater.

The clustering result of the six selected variables of air preheater operating data is shown in Figure 3. Those that need to be addressed include, we select the parameter d_c of DPC as 5%, and PCA is adopted to approve the clustering ability and to visualize the clustering results. Although we use PCA to project the data into two-dimensional space, the boundary of the classification result is still ambiguous. In order to distinguish the differences between the four clusters of data, the result of ΔP is shown in Figure 4 with different colors. Those green data points indicate a good state with relatively low air leakage during this period of operation and we name it as state one. Cyan ones named as state two represent a slightly inferior state, whose leakage is slightly greater than state one. The rest points with the color of blue and red are named as state three and state four, respectively. State four has the maximum air leakage during this period of operation, state four is a little bit better than state four. Therefore, we can label these data to construct the state library for state detection in the next step.

4.1.1 EKNN based state detection.

After clustering processing using DPC, the historical data is classified and labelled as a state library. The following operating data then can be identified by EKNN based on the state library. Nevertheless, according to Figure 3, the classification boundary is uncertain due to the relatively high-dimensional data. In addition, noises contained in the operating data makes the classification problem more difficult. To validate the classification results, two popular criteria Accuracy and normalization mutual information(NMI)[16] are used to evaluate the effectiveness. And we define that class denotes the ground truth and cluster denotes the results produced by a specific algorithm.

According to Figure 1, in this step, operating data can be identified by EKNN based on the labeled historical data. In this paper, two simulation methods are used to verify the validity of EKNN. First, according to the time sequence, the data is divided into two parts. The continuous 500 minutes data are used for testing, and the rest 4500 minutes data are for constructing the label library. The second method is to randomly sample 10% from the data for testing, and the remaining for constructing the label library. Both test methods are repeated ten times and the results are summarized in Table 1.

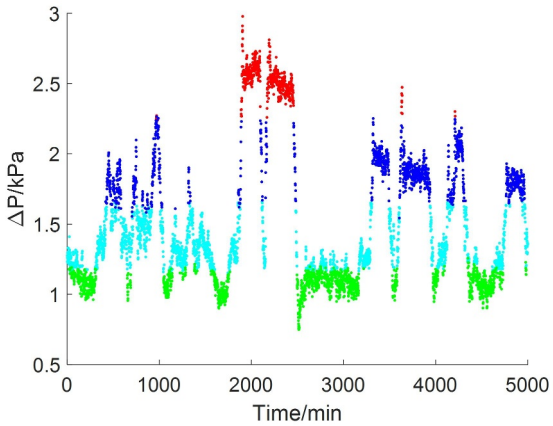

Fig. 4. ΔP of operating data for air preheater.

Table 1 shows that in both tests EKNN has achieved good results for state detection. Accuracy and NMI are a bit lower in continuous sampling test, the reason might be that continuous sampling deteriorate the integrality of label library in this finite sample case.

Table 1. Test results for state detection

Test method	Accuracy	NMI[16]
Continuous sampling	0.9574 ± 0.0351	0.8288 ± 0.1139
Random sampling	0.9734 ± 0.0039	0.9028 ± 0.0103

5 Conclusion

In this paper, we proposed a data-driven state monitoring method for the air preheater system of thermal power plant. By adopting DPC to process historical data, label library can be constructed based on the true existing states in the data and almost little expertise is required. Then, EKNN improve the classification performance in the presence of imprecise and ambiguous information. The state monitoring effectiveness of proposed method is verified in two test method and acquired good results in terms of accuracy and NMI.

Acknowledgments

The authors would like to acknowledge the National Natural Science Foundation of China (NSFC) under Grant 51506041, 51506029, the Natural Science Foundation of Jiangsu Province, China under Grant BK20150631.

References

1. Chen, Xiao Long, et al. "Evidential KNN-Based Condition Monitoring and Early Warning Method with Applications in Power Plant." Neurocomputing (2018).
2. Widodo, Achmad, and B. S. Yang. "Support vector machine in machine condition monitoring and fault diagnosis." Mechanical Systems & Signal Processing 21.6(2008):2560-2574.
3. Haykin, Simon. Neural Networks: A Comprehensive Foundation (3rd Edition). Macmillan, 1998.
4. Wang, Wilson Q., M. F. Golnaraghi, and F. Ismail. "Prognosis of machine health condition using neuro-fuzzy systems." Mechanical Systems & Signal Processing 18.4(2004):813-831.
5. Sang, Wook Choi, et al. "Adaptive Multivariate Statistical Process Control for Monitoring Time-Varying Processes." Industrial & Engineering Chemistry Research 45.9(2006):687-706.
6. Rodriguez, Alex, and A. Laio. "Clustering by fast search and find of density peaks." Science 344.6191(2014):1492.
7. Denoeux, T. "A k-nearest neighbor classification rule based on Dempster-Shafer theory." Systems Man & Cybernetics IEEE Transactions on 25.5(2008):804-813.
8. Du, Mingjing, S. Ding, and H. Jia. "Study on density peaks clustering based on k-nearest neighbors and principal component analysis." Knowledge-Based Systems 99(2016):135-145.
9. Keller, J. M., M. R. Gray, and J. A. Givens. "A fuzzy K-nearest neighbor algorithm." IEEE Transactions on Systems Man & Cybernetics SMC-15.4(2012):580-585.
10. RONALD R. YAGER. "DECISION MAKING UNDER DEMPSTER-SHAFER UNCERTAINTIES." International Journals of General Systems20.3(2008):233-245.
11. Chew. "Rotary air preheaters on power-station boilers." (1985).
12. Skiepko, T. "Experimental results concerning seal clearances in some rotary heat exchangers." Heat Recovery Systems & Chp 8.6(1988):577-581.
13. Shah, R. K., and T. Skiepko. "Influence of leakage distribution on the thermal performance of a rotary regenerator." Applied Thermal Engineering 19.7(1999):685-705.
14. Jestin, Louis, W. Fuls, and M. Pronobis. "A numerical study of air preheater leakage." Energy 92(2015):87-99.
15. Cai, Mingkun, et al. "A study on the direct leakage of rotary air preheater with multiple seals." Applied Thermal Engineering 59.1-2(2013):576-586.
16. Vinh, Nguyen Xuan, J. Epps, and J. Bailey. "Information theoretic measures for clusterings comparison: is a correction for chance necessary?." International Conference on Machine Learning ACM, 2009:1073-1080.