Kernel Based Manifold Learning for Complex Industry Fault Detection

Jian Cheng and Yi-nan Guo

CUMT-IoT Perception Mine Research Center,
School of Information and Electrical Engineering,
China University of mining and Technology, Xuzhou, 221116, China
chengjian@cumt.edu.cn

Abstract. Accurate and rapid fault detection based on the data from industry process is very important for the process control. This paper introduces a new multivariate statistical process control approach for fault detection using kernel method based manifold learning algorithm combining T^2 statistic. The proposed approach is effective in the fault detection, which has two stages. Stage I: a kernel method based locally linear embedding is employed to extract the nonlinear features, preserve local structure and reduce dimensionality of the multivariate input data, and a new low-dimensional embedding method is developed to solve the "out-of-sample" problem. Stage II: the fault detection is performed by T^2 statistic with control limits derived from the eigenanalysis of the kernel matrix in the Hilbert feature space. In this study, the method is applied for the fault detection of the benchmark Tennessee Eastman (TE) challenge process. The proposed method has been compared with conventional methods in terms of performances such as detection accuracy, detection delay and false alarm rate. It is demonstrated that the proposed method outperformed the others in fault detection on TE process.

Keywords: Manifold learning, Kernel method, Fault detection, TE process.

1 Introduction

Generally, process monitoring consists of measuring and controlling several process variables simultaneously[1], which is increasing difficult to detect and diagnose the fault states if multiple process variables exhibit outliers or deviations at the same time. To overcome this disadvantage, multivariate quality control methods has been employed by monitoring the interactions of several process variables and determining hidden factors using dimensionality reduction methods[2], such as principal component analysis (PCA), dynamic principal component analysis (DPCA), Fisher Discriminant Analysis (FDA) and partial least squares (PLS)[3][4]. Subsequently, applying multivariate statistics to the low-dimensional data representations produced by these methods, faults can be

H. Yin et al. (Eds.): IDEAL 2013, LNCS 8206, pp. 392–400, 2013.

detected with greater accurate. However, for some complex situations in industrial field with particularly nonlinear characteristics, there are poor performances due to the assumption that the process data are linear for these methods.

This shortcoming of linear methods has led to the development of nonlinear methods. Then a series of nonlinear PCA method based on neural networks has been developed[5][6]. However, it is difficult to determine structure of neural network and neural network training consumes a lot of time. Another method is based on kernel method, for example, kernel PCA (KPCA)[7] and one-class support vector machines (1-class SVM)[20]. Although these methods can deal with the nonlinear problem, they all do not consider the potential structure of the input data.

For nonlinear dimensionality reduction, Manifold learning is a perfect tool which can discover and retain the structure of high dimensional data sets for a better understanding of the data. Several different manifold learning algorithms have been developed to perform dimensionality reduction, such as Isomap[9], locally linear embedding (LLE)[10], Laplacian eigenmaps (LE)[11], and Maximum Variance Unfolding (MVU)[12], etc.

The LLE algorithm is considered one of effective manifold learning algorithms for dimensionality reduction, and has been used to solve various problems in information processing, pattern recognition and data mining[13][14]. LLE algorithm computes a different local quantity, and calculates the best coefficients to approximate each point by a weighted linear combination of its neighbors, and then tries to find a set of low-dimensional points, which can be linearly approximated by its neighbors with the same coefficients that have been determined from high-dimensional points. However, LLE may fail to deliver good performance when the data structure is nonlinear. Moreover, it faces the difficulty of how to implement the map on new testing data points.

Because the the sensor data collected from process is typically nonlinear, high-dimensional and generally correlated, it is necessary to develop a new method to tackle the fore-mentioned drawbacks of traditional methods. Recently, there has been great interest in developing low dimensional representations through kernel based methods[15][16][17]. These methods can efficiently discover the nonlinear structure of data and evaluate the map on out-of-sample data. In this paper, the proposed method includes two stages. Firstly, we present a kernel method based LLE algorithm (KLLE) to reduce the dimensionality of the input data and obtain a low-dimensional embedding data. In the second stage, we adopt Hotelling's T^2 statistic chart[1] in the embedding data to determine the control limit and detect the fault state. This method can not only deal with the nonlinear problem, but also preserve the potential structure in the input data points.

The remainder of the paper is organized as follows. Section 2 presents kernel method based manifold learning. In Section 3, we introduces T^2 statistic in Hilbert feature space for fault detection. We experimentally evaluate the performance of our proposed method using TE process data in section 4. Section 5 concludes this paper.

2 Kernel Method Based Manifold Learning

2.1 Kernel Method Based LLE

Let $X = \{x_1, x_2, \cdots, x_n\}$ be a given data set of n points in $\mathbb{R}^m$, sampled from a $d-$dimensional manifold ($d \leq m$). LLE constructs a dataset $Y = \{y_1, y_2, \cdots, y_n\} \subseteq \mathbb{R}^d$. There are the details of LLE in[10]. With the kernel trick, suppose that the input space X is mapped into a Hilbert feature space $\mathbb{H}$ through a nonlinear mapping function $\phi : \mathbb{R}^m \to \mathbb{H}$[15]. In Hilbert feature space $\mathbb{H}$, the nearest neighborhood of $\phi(x_i)$ is $\{\phi(x_i^j), j = 1, \cdots, l\}$, where l is the number of nearest neighbors of $\phi(x_i)$. As in LLE, it is worth to note that the Euclidean distance between two data points in the Hilbert feature space can be computed according to

$$\| \phi(x_i) - \phi(x_j) \|^2 = k(x_i, x_i) + k(x_j, x_j) - 2k(x_i, x_j). \tag{1}$$

Then, in Hilbert feature space, the first step in LLE is to determine the neighbor set for $\phi(x_i)$ and to learn the local linear structure of the neighbor set by solving

$$\min \| \phi(x_i) - \sum_{j=1}^{l} w_{ij} \phi(x_i^j) \|^2, \tag{2}$$

The weights w_{ij} can be computed by minimize Equation 2 with the constraints $\sum_{j=1}^{l} w_{ij} = 1$. We can get the following Lagrange formulation

$$L(W_i) = W_i^T C_i W_i - \lambda(W_i^T \mathbf{e} - 1), \tag{3}$$

where C_i is the local kernel matrix of $\phi(x_i)$ in $\mathbb{H}$, and

$$\begin{aligned} C_i(j,k) &= (\phi(x_i) - \phi(x_i^j))^T (\phi(x_i) - \phi(x_i^k)) \\ &= k(x_i, x_i) - k(x_i, x_i^k) - k(x_i, x_i^j) + k(x_i^j, x_i^k). \end{aligned} \tag{4}$$

Equation 3 which is subjected to $W_i^T \mathbf{e} = 1$ ($\mathbf{e}$ is a one dimension vector consisting of ones) has the closed form solution $W_i = C_i^{-1}\mathbf{e}/(\mathbf{e}^T C_i^{-1}\mathbf{e})$ [18]. C_i is a positive definite matrix, the eigen-decomposition of C_i is of form $C_i = U^T \Lambda U$, then

$$W_i = (U^T \Lambda^{-1} U\mathbf{e})/(\mathbf{e}^T U^T \Lambda^{-1} U\mathbf{e}). \tag{5}$$

Hence, the reconstruction weights W are computed by C's eigenvalues and eigenvectors.

2.2 KLLE Embedding

In Hilbert feature space, we suppose that the embedding Y can be given by $Y = \Gamma^T \phi(X)$, where Γ is a linear transformation matrix and $y_i = \Gamma^T \phi(x_i) \in \mathbb{R}^d$.

Now, we turn to the problem of finding a transformation matrix Γ in Hilbert feature space $\mathbb{H}$. The best low-dimensional embedding Y can be computed by

$$\begin{aligned}\sum_i \parallel y_i - \sum_j w_{ij} y_j \parallel^2 &= \parallel Y(I-W) \parallel^2 \\ &= tr[\Gamma^T \phi(X) \widetilde{M} \phi(X)^T \Gamma],\end{aligned} \tag{6}$$

where $Y = [y_1, y_2, \cdots, y_n] = \Gamma^T \phi(X)$ and $\widetilde{M} = (I-W)(I-W)^T$. Since each column of Γ should lie in the span of $\phi(x_1), \phi(x_2), \cdots, \phi(x_n)$, we can write

$$\Gamma = [\sum_i a_1(i)\phi(x_i), \sum_i a_2(i)\phi(x_i), \cdots, \sum_i a_d(i)\phi(x_i)] = \phi(X)A, \tag{7}$$

where $a_k(i), k = 1, \cdots, d$, denotes the ith entry of the coefficient vector a_k, and $A = [a_1, a_2, \cdots, a_d] \in \mathbb{R}^{n \times d}$. Substituting Equation 7 to Equation 6, we can obtain

$$\min_{\Gamma} \sum_i \parallel y_i - \sum_j w_{ij} y_j \parallel^2 = \min_A tr(A^T K \widetilde{M} K A), \tag{8}$$

where $K = \phi(X)^T \phi(X)$ is a $(n \times n)$ symmetric kernel matrix whose entries are $K(i,j) = k(x_i, x_j)$. Similarly, $\frac{1}{n} YY^T = I$ becomes to $\frac{1}{n} A^T K K A = I$.

Finally, the constrained minimization problem above is converted to the following generalized eigenvalue problem

$$K \widetilde{M} K a = \lambda K K a. \tag{9}$$

And the matrix A is determined by the eigenvectors corresponding to the bottom d nonzero eigenvalues of Equation 9. Once A is obtained, for any data point x in high dimensional space $\mathbb{R}^m$, it can be mapped to a low dimensional space point $y \in \mathbb{R}^d$ by

$$y = \Gamma^T \phi(x) = A^T [k(x_1, x), k(x_2, x), \cdots, k(x_n, x)]^T, \tag{10}$$

and we can obtain another form as follow

$$y_j = \sum_i^n a_j^i k(x_i, x), \quad j = 1, \cdots, d, \tag{11}$$

where y_j is the jth entry of embedding coordinate y.

3 T^2 Statistic for Fault Detection

Using the KLLE, we can collect the embedding coordinates $Y = [y_1, y_2, \cdots, y_n]^T \in \mathbb{R}^{n \times d}$. A measure of the variation within the KLLE is given by Hotelling's T^2 statistic[2]. T^2 is the sum of the normalized squared scores, and is defined as

$$T^2 = Y \Lambda^{-1} Y^T, \tag{12}$$

where Λ^{-1} is the diagonal matrix of the inverse of the eigenvalues associated with the embedding coordinates obtained by the generalized eigen decomposition of Equation 9.

Given a level of significance α, appropriate threshold value for the T^2 statistic can be determined automatically by applying the probability distributions. In this study, the control threshold for T^2 is obtained using the $F-$distribution

$$T^2_{d,n,\alpha} \sim \frac{d(n-1)}{n-d} F_{d,n-d,\alpha}, \tag{13}$$

where n is the number of samples and d is the number of embedding dimensions.

4 Experiments and Results

4.1 Tennessee Eastman Process and Data Acquisition

The Tennessee Eastman (TE) plant-wide industrial process control problem was proposed by Downs and Vogel, described in detail in[19][3]. The dataset includes all the manipulated and measured variables, except the agitation speed of the reactor's stirrer for a total of $m = 52$ observation variables(The agitation speed was not included because it was not manipulated).

The simulation time for each run is 96 hours. The first 48 hours are operated under normal operating conditions, the fault is induced after 48 hours. A sampling interval of 3 minutes is used to collect the simulated data for the training (normal state) and testing data (fault state). The total number of observations generated for each run is 1920 samples. The normal operating training data consists of 960 samples. When no faults induced, we can obtain other 960 normal state samples as validation data. The control threshold of T^2 statistic can be set based on this validation data in the next section. In the following section, the performance of KLLE is compared with PCA, DPCA, KPCA and 1-class SVM for TE process.

4.2 Fault Detection of the TE Process

We compared the fault detection performance of kernel based LLE (KLLE) monitoring method with that of PCA[3], DPCA[4], KPCA and 1-class SVM[8] based monitoring method. Fault detection performance was evaluated by detection accuracy and false alarm rate of each method. In accordance with the works by Mahadevan et al.[8] and Chiang et al.[3], the fault is indicated only when six consecutive statistic values exceed the threshold and the detection delay of a monitoring chart is defined as the time gap between the introducing of fault and the statistic value exceeding its upper control threshold for the first time. Since it is unfair to compare detection accuracy and detection delays of all methods when they have different false alarm rate, in computing above indices, the control threshold for each monitoring statistic in each method was adjusted to the 10th highest value of the normal operating validation data. In this way, the adjusted threshold corresponds to the 99% confidence limit.

Table 1. The detection accuracy for the testing data

Model	Fault1	Fault2	Fault4	Fault5	Fault6	Fault7	Fault8	Fault10	Fault11
PCA+T^2	99.1	97.3	-	23.5	98.9	91.3	96.3	33.3	20.5
DPCA+T^2	99.4	98.1	-	24.2	98.7	84.1	97.2	42.0	19.9
KPCA+T^2	**100**	99.3	83.7	29.8	**100**	**100**	97.6	80.8	81.3
1-class SVM	99.8	98.6	**99.6**	**100**	**100**	**100**	97.9	87.6	69.8
KLLE+T^2	**100**	**99.6**	96.2	**100**	**100**	**100**	**99.4**	**96.6**	**87.5**

Model	Fault12	Fault13	Fault14	Fault16	Fault17	Fault18	Fault19	Fault20	Fault21
PCA+T^2	97.9	94.0	84.3	16.7	75.1	88.7	-	29.7	26.4
DPCA+T^2	99.0	95.1	93.9	21.7	76.0	88.9	-	35.6	35.6
KPCA+T^2	98.4	95.5	**100**	77.6	95.2	91.3	75.7	72.2	81.7
1-class SVM	**99.9**	95.5	**100**	89.8	**95.3**	90.0	83.9	90.0	52.8
KLLE+T^2	98.8	**97.5**	**100**	**89.9**	87.9	**92.9**	**84.4**	**96.6**	**81.9**

"-" denotes that the fault cannot be detected.

According to Chiang et al.[3] and Russell et al.[4] detection accuracies for faults 3, 4, 9, 15 and 19 were very low because there was no observable change in their means, variances or the higher moments. Hence it has been considered that these faults are unobservable and cannot be detected by any traditional statistical technologies. In the work of Mahadevan et al.[8], 1-class SVM method was also not able to detect faults 3, 9 and 15 efficiently, but can detect faults 4 and 19. Therefore, in this paper, our comparative experiments have not included faults 3, 9 and 15, although we can monitor all faults in TE process by the proposed method.

The fault detection accuracies are calculated and tabulated in Table 1. The maximum fault detection value obtained for each of the faults has been highlighted in bold face. As expected, DPCA statistic have performed better than that of PCA for most faults, this indicates the potential advantage of taking serial correlation into account by DPCA when developing fault detecting procedures. However, the fault detection accuracies of PCA and KPCA are still low, and the average of detection accuracies of PCA and DPCA over all faults are only 67.1% and 69.3% respectively, because they are linear method which cannot capture and model the nonlinear of TE process.

Table 2. The average of detection accuracy (ADA), detection delays (ADD) and false alarm rate (AFAR)

Model	ADA(%)	ADD	AFAR(%)
PCA+T^2	67.1	102.0	1.60
DPCA+T^2	69.3	84.4	**0.99**
KPCA+T^2	87.0	26.2	1.47
1-class SVM	91.7	32.8	1.36
KLLE+T^2	**94.9**	**1.4**	1.25

Obviously, the fault detection accuracies based on KPCA, 1-class SVM and KLLE are much higher than PCA and DPCA due to their nonlinearity via kernel method. For some faults such as fault 1, 2, 6, 7 and 14, these kernel based

method have similar detection accuracy, but 1-class SVM and KLLE are better than KPCA for most fault states demonstrated by average detection accuracy in Table 2. It is also observed that KLLE performs better than 1-class SVM for most of the faults except fault 4, 12 and 17, and the average of detection accuracies of 1-class SVM and KLLE are up to 91.7% and 94.9% in Table 2 respectively. In the case of fault 5, KLLE easily distinguishes the faulty data from the normal operating data. However, KPCA monitoring chart cannot detect it.

The false alarm rate of all the methods are summarized in Table 2. It can be seen that the DPCA method is much better than the rest of the models. Nevertheless, the false alarm rates of all the methods are well within acceptable limits. However, it should be noted that KLLE performs much better than other methods in its reduced detection delay and increased fault detection accuracy.

Mentioned above, the faults 3, 9 and 15 are unobservable from the testing data according to PCA, DPCA, KPCA and 1-class SVM. Although these three fault states have no observed change in the mean, the variance and the higher order variance, they can induce the change of data structure captured by KLLE. Therefore, it is encouraging that these three faults can be successfully detected by our proposed method. The detection accuracies of the faults 3, 9 and 15 are 77.7%, 90.7% and 73.0%, and the false alarm rates are 2.50%, 1.35% and 1.88% respectively.

From the results above, it is obviously shown that KLLE performs excellently on fault detection. Two facts demonstrate this capability. On the one hand, in nonlinear structures data set, KLLE preserves intrinsic properties more than the PCA, DPCA and KPCA. On the other hand, KLLE preserves the data structure more than 1-class SVM, and we also found that the time consumption of KLLE is smaller than 1-class SVM.

5 Conclusion

The application of machine learning to data mining and analysis in area of TE process is rapidly gaining interest in the community. In this paper, we presented a new effective approach to detect all the fault states in TE process. Different from conventional monitoring methods, KLLE can not only capture and model the nonlinearity, but also preserve the potential structure in the data. Moreover, the proposed method can get a implicit mapping relation between the original data space and the low-dimensional feature space by using kernel trick and linear projection, which make it possible to monitor online. For the fault detection of TE process, KLLE performed much better than other conventional monitoring methods such as PCA, DPCA, KPCA and 1-class SVM. Compared with these technologies, the proposed method could detect the fault with an increased detection accuracy and a considerable reduction in detection delay, and could generalize well to all the faults of TE process. Meanwhile, the false alarm rates were also within the acceptable limits, thus making it more useful and feasible for industrial online application.

Acknowledgments. This work is partially supported by the Fundamental Research Funds for the Central Universities (Grant No. 2011QNB24) and Jiangsu Natural Science Foundation (Grant No. BK2010183). The authors would like to thank the anonymous reviewers and editors for their help comments and suggestions.

References

1. Montgomery, D.C.: Introduction to Statistical Quality Control. John Wiley & Sons (2005)
2. Yang, K., Trewn, J.: Multivariate Statistical Methods in Quality Management. McGraw-Hill Professional (2004)
3. Chiang, L.H., Russell, E.L., Braatz, R.D.: Fault Detection and Diagnosis in Industrial Systems. Springer (2001)
4. Russell, E.L., Chiang, L.H., Braatz, R.D.: Fault detection in industrial processes using canonical variate analysis and dynamic principal component analysis. Chemometrics and Intelligent Laboratory Systems 51(1), 81–93 (2000)
5. Jia, F., Martin, E.B., Morris, A.J.: Nonlinear principal components analysis with application to process fault detection. Journal of Systems Science 31(5), 1473–1487 (2001)
6. Dong, D., McAvoy, T.J.: Nonlinear principal component analysis based on principal curves and neural networks. Computers and Chemical Engineering 20(1), 65–78 (1996)
7. Schölkpof, B., Smola, A., Müller, K.-R.: Nonlinear Component Analysis as a Kernel Eigenvalue Problem. Neural Computation 10(5), 1299–1319 (1998)
8. Mahadevan, S., Shah, S.L.: Fault detection and diagnosis in process data using one-class support vector machines. Journal of Process Control 19(10), 1627–1639 (2009)
9. Tenenbaum, J.B., Silva, V., Langford, J.C.: A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science 290, 2319–2323 (2000)
10. Roweis, S.T., Saul, L.K.: Nonlinear Dimensionality Reduction by Locally Linear Embedding, vol. 290, pp. 2323–2326 (2000)
11. Belkin, M., Niyogi, P.: Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. Neural Computation 15, 1373–1396 (2003)
12. Weinberger, K.Q., Saul, L.K.: An Introduction to Nonlinear Dimensionality Reduction by Maximum Variance Unfolding. In: Proceedings of the 27th National Conference on Artificial Intelligence, pp. 1683–1686. AAAI Press (2001)
13. Elgammal, A.M., Lee, C.S.: Separating style and content on a nonlinear manifold. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern, pp. 478–485. IEEE Press (2004)
14. Mekuz, N., Bauckhage, C., Tsotsos, J.K.: Face recognition with weighted locally linear embedding. In: Proceedings of The Second Canadian Conference on Computer and Robot Vision, pp. 290–296. IEEE Press (2005)
15. Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press (2004)
16. Choi, H., Choi, S.: Robust Kernel Isomap. Pattern Recognition 40(3), 853–862 (2007)
17. Yu, X., Wang, X., Liu, B.: Supervised kernel neighborhood preserving projections for radar target recognition. Signal Processing 88(9), 2335–2339 (2008)

18. Saul, L.K., Roweis, S.T.: Think globally, fit locally: unsupervised learning of low dimensional manifolds. The Journal of Machine Learning Research 4, 119–155 (2003)
19. Downs, J.J., Vogel, E.F.: A Plant-wide Industrial Process Control Problem. Computers & Chemical Engineering 17(3), 245–255 (1993)
20. Schölkopf, B., Platt, J., Shawe-Taylor, J., Smola, A., Williamson, R.: Estimating the support of a high-dimensional distribution. Nerual Computation 13(7), 1443–1471 (2001)