18

Spectral Clustering and Its Application in Machine Failure Prognosis

Weihua Li[1,2], Yan Chen[2], Wen Liu[1] and Jay Lee[2]
[1]*School of Mech. & Auto. Eng,*
Souch China University of Technology,
[2]*NSFI/UCRC Center for Intelligent Maintenance System,*
University of Cincinnati,
[1]*China*
[2]*USA*

1. Introduction

Machine fault prognosis and health management has received intensive studies for several decades, and various approaches have been taken, such as statistical signal processing, time-frequency analysis, wavelet, and neural networks. Among of them, pattern recognition method provides a systematic approach to acquiring knowledge from fault samples. In fact, mechanical fault diagnosis is essentially a problem of pattern classification.

Many pattern recognition methods have been studied and applied in machine condition monitoring and fault prognosis. Campbell proposed a linear programming approach to engine failure detection (Campbell&Bennett, 2001). In Ypma's study, different learning methods, such as Independent Component Analysis, Self Organising Map, and Hidden Markov Models, were applied in fault feature extraction, novelty detection and dynamic fault recognition (Ypma, 2001). Ge et.al (2004)proposed a support vector machine based method for sheet metal stamping monitoring. Harkat et.al(2007) applied non-linear principal component analysis in sensor fault detection and isolation. Lei and Zuo (2009) implemented the Weighted *k* Nearest Neighbour algorithm to identify the gear crack level.

However, the information of machine incipient fault is always weak and contaminated by strong noises, and there is always lack of fault samples to train the learning machine. Therefore, the key issue is how to select sensitive features from the dataset for machine incipient faults prognosis, which is related to feature selection and dimension reduction, and is very useful for fault classification.

In most of medical and clinic applications, when the dimensionality of the data is high, for reducing computation complexity, some techniques might be used to project or embed the data into a lower dimensional space while retaining as much information as possible. Classical linear examples are Principal Component Analysis (PCA) (Jolliffe.2002) and Multi-Dimensional Scaling (MDS) (T. F. Cox & M. A. Cox, 2001). The coordinates of the

data points in the lower dimension space might be used as features or simply a mean to visualize the data.

However, for common PHM(Prognostic and Health Management) applications, the dimensionality of the data is not as high as those in medical research, and the mapping techniques are mainly applied to reveal the correlation of features as to increase the accuracy of fault detection and identification. The selection of features also can avoid unnecessary sensors used in machine monitoring, considering the high cost maintaining. Nomikos and MacGregor(1994) firstly presented a PCA approach for monitoring batch process, the history information was linear projected onto a low-dimensional space that summarized the key characteristics of normal behaviour by both variable and their time histories. Considering that minor component discarded in PCA might contain important information on nonlinearities, a large amount of nonlinear methods were presented for the process monitoring and chemical process modelling (Dong & McAvoy,1996; Kaspar & Ray,1992; Sang et. al,2005), such as Kernal PCA (Schölkopf,1998).

Non-linear dimensionality mapping methods are more frequently recognized as non-linear manifold learning methods. The manifold learning is the process of estimating a low-dimensional underlying structure embedded in a collection of high-dimensional data(Tenenbaum et. al, 2000; Roweis & Saul, 2000). Instead of using Euclidian distance to measure samples' similarity in input space, samples' similarity in latent space is measured by their geodesic or short path distance. The deceptive close distance in the high-dimensional input space can be corrected.

Spectral clustering is a graph-theory-based manifold learning method, which can be used to dissect the graph and get the clusters for exploratory data analysis. Compared with the traditional algorithms such as k-means, spectral clustering has many fundamental advantages. It is more flexible, capturing a wider range of geometries, and it is very simple to implement and can be solved efficiently by standard linear algebra methods. It has been successfully deployed in numerous applications in areas such as computer vision, speech recognition, and robotics. Moreover, there is a substantial theoretical literature supporting spectral clustering (Kannan et.al,2004; Luxburg,2007,2008).

In most PHM applications, multi-groups of data sets from different failure modes are frequently nonlinearly distributed and mixed in a high dimensional feature space. However, an "unfolded" feature space is expected as to differentiate these degradation patterns by a designed classifier.

In this part, we first propose a spectral clustering based feature selection method used for machine fault feature extraction and evaluation, and then the samples with selected features are input into a density-adjustable spectral kernel based transductive support vector machine to train and to get the prognosis results.

2. Spectral clustering feature selection

2.1 Basics of graph theory

Given a d-dimentsional data points $\{x_1, \ldots, x_n\}$, and the similarity between all pairs of data points x_i and x_j is noted as w_{ij}. According to graph theory, the data points can be represented

by an undirected data graph G=(V,E). Each node in this graph represents a data point x_i. Two nodes are connected if the similarity w_{ij} between the corresponding data x_i and x_j is positive or larger than a certain threshold, and the edge is weighted by w_{ij}. These data points can be divided into several groups such that points in the same group are similar and points in different groups are dissimilar to each other.

2.2 Laplacian embedding

BelKin(2003) indicated that Laplacian Eigenmaps used spectral techniques to perform dimensionality reduction. This technique relies on the basic assumption that the data lies in a low dimensional manifold in a high dimensional space. The Laplacian of the graph obtained from the data points may be viewed as an approximation to the Laplace-Beltrami operator defined on the manifold. The embedding maps for the data come from approximations to a natural map that is defined on the entire manifold.

The popular Laplacian Embedding algorithm includes the following steps, as shown in Fig.1.

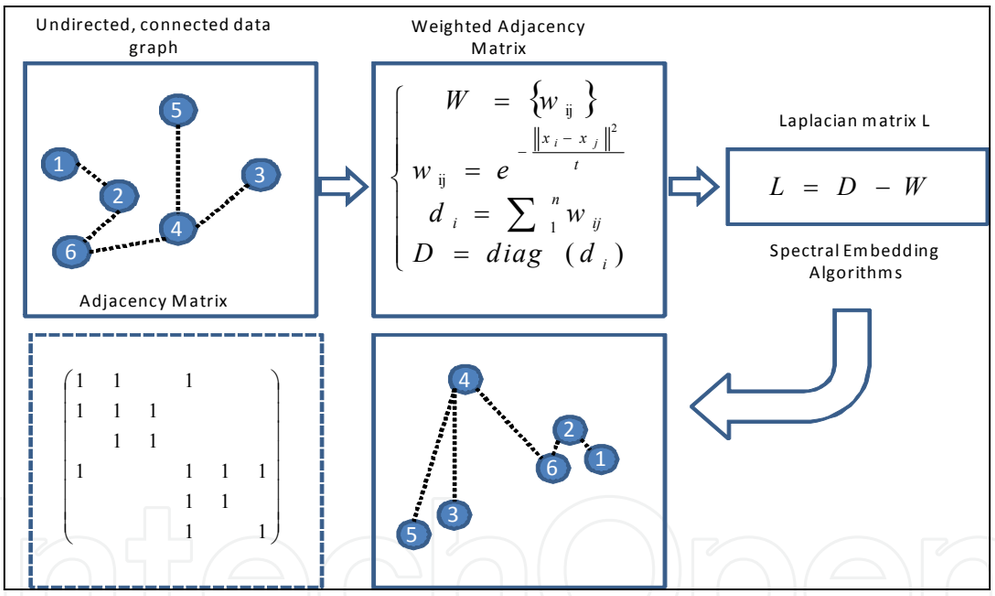


Fig. 1. The procedure of Laplacian Embedding Algorithm

Step 1: The d-dimensional dataset is viewed as an undirected data graph [10] , G = (V, E) with node set V={x_1,...,x_n}. Every node in the graph is one point in $\Re_d$. An edge is used to link node i and node j, if they are close as ε–neighborhoods which means the distance between nodes $\mathbf{X}_i$ and $\mathbf{X}_j$ satisfying $\|\mathbf{X}_i - \mathbf{X}_j\| < \varepsilon$, or if node $\mathbf{X}_i$ is among n nearest neighbors of $\mathbf{X}_j$ or $\mathbf{X}_j$ is among n nearest neighbors of $\mathbf{X}_i$.

Step 2: Each edge between two nodes $\mathbf{X}_i$ and $\mathbf{X}_j$ carries a non-negative weight $w_{ij}\geq 0$. The weighted adjacency matrix of the graph is the matrix $\mathbf{W} = \{w_{ij}\}, i, j = 1,...,n$. There are different methods to configure the weight matrix. For example, the most common is

$$w_{ij} = \begin{cases} 1 & \text{if } x_i \text{ and } x_j \text{ is connected} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Or Heat kernel

$$w_{ij} = \begin{cases} e^{-\|x_i - x_j\|^2 / t} & \text{if } x_i \text{ and } x_j \text{ is connected} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

The degree of a node $X_i \in V$ is defined as $d_i = \sum_1^n w_{ij}$. The degree matrix D is defined as the diagonal matrix with $\{d_1, d_2,\ldots,d_n\}$ on its diagonal. The un-normalized graph Laplacian matrix is defined by Luxburg(2007) as: **L = D - W** .

Step 3: The Laplacian Eigenmap (on normalized Laplacian matrix) is computed by spectral decomposition for eigenvectors problem of $\mathbf{L}y = \lambda \mathbf{D}y$. The image of X_i under the embedding is converted into the lower dimensional space $\Re m$, given by ordered eigenvalues: $\{y_1(i), y_2(i),..., y_m(i)\}$. This decomposition provides significant information about the graph and distribution of all points. It has been proven experimentally that the inner natural groups of dataset are recovered by mapping the original dataset into the space spanned by eigenvectors of the Laplacian matrix(Belkin & Niyogi,2003).

2.3 Supervised feature selection criterion by Laplacian scores

Given a graph G, the Laplacian matrix **L** of G is a linear operator on any feature vector from $\mathbf{f} = \{\mathbf{f}_1,...,\mathbf{f}_m\}, \mathbf{f}_i \in R^n$

$$\mathbf{f}_k{}^T \mathbf{L} \mathbf{f}_k = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} (x_{ki} - x_{kj})^2 \quad (3)$$

The equation quantifies how much the feature vector is consistent with the structure of the G locally. For the instances closer to each other, the features that have similar value for them are contributes more on the dissimilarity matrix that is consistent with data structure. The flatter the feature value is over all instances, the smaller the value of the equation. However, instead of the feature consistency only considering instances with small distance, a complete definition of feature consistency with the data structure is clarified as the following:

Definition 1: (feature local consistency with data graph)

Given data graph G= (**V**, **E**) ($V=\{X_1,...,X_n\}, E=\{W_{ij}\}$), the feature **f** is a locally consistent variant of G at level h (0<h<1) for a clustering C over G. If for every cluster C_k of C, there is

$$\frac{\frac{1}{n_k} \sum_{i,j \in C_k} (\mathbf{f}_i - \mathbf{f}_j)^2}{\frac{1}{n_k (n - n_k)} \sum_{i \in C_k, j \notin C_k} (\mathbf{f}_i - \mathbf{f}_j)^2} = h_k \quad (4)$$

And $h = \max(h_k)$ is defined as feature consistency index.

The definition is a ratio between inner and intra cluster variation caused by the individual feature. Perfect clustering expects less variance inter-cluster and the inverse for intra-clusters. If the feature **f** contributes to better clustering, the nominator tends to be smaller and denominator is larger. Therefore h_k is expected to be smaller. The feature consistency index *h* indicates the features's weakest separablility for clustering C

In terms of graph theory, similar criterion can be formulated based on Eq.(4), and configure data graph G with following similarity measurement,

$$w^{(1)}{}_{ij} = \begin{cases} 1/n_k & i, j \in C_k \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$w^{(2)}{}_{ij} = \begin{cases} 0 & i, j \in C_k \\ -1/(n - n_k) & otherwise \end{cases} \tag{6}$$

Where $w^{(1)}{}_{ij}$ is the similarity measurement of samples within-class, and $w^{(2)}{}_{ij}$ that of samples between-class. Then the sequence of instances can be reordered to make the adjacency matrix carry closer instances along its diagonal.

$$\mathbf{W}^{(2)} = \begin{pmatrix} W_1{}^{(2)} & \cdots & -1/n_1(n - n_1) \\ \vdots & \ddots & \vdots \\ 1/n_p(n - n_p) & \cdots & W_{n_p}{}^{(2)} \end{pmatrix} \tag{7}$$

$$\mathbf{W}^{(1)} = \begin{pmatrix} W_1{}^{(1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & W_{n_p}{}^{(1)} \end{pmatrix} \tag{8}$$

As proved by He et.al (2006), Laplacian score of *r*-th feature is as the follows:

$$L_r = \frac{\tilde{f}_r{}^T L \tilde{f}_r}{\tilde{f}_r{}^T D \tilde{f}_r} \tag{9}$$

Because of $\tilde{\mathbf{f}}_r{}^T \mathbf{L} \tilde{\mathbf{f}}_r = \mathbf{f}_r{}^T \mathbf{L} \mathbf{f}_r$ (He et.al, 2006), and with the weight matrix as $\mathbf{W}^{(2)}$ and $\mathbf{W}^{(1)}$ as well as their degree diagonal matrixes,

$$\mathbf{D}^{(1)} = -\mathbf{D}^{(2)} = \begin{pmatrix} 1/n_1 & 0 & 0 \\ 0 & \ddots & \vdots \\ 0 & \cdots & 1/n_p \end{pmatrix} \tag{10}$$

The two Laplacian score $\mathbf{L}_r^{(1)}$ and $\mathbf{L}_r^{(2)}$ have same absolute value of denominators. If there exists clustering C={C_1,…, C_p } over data graph G, the nominators are as the following

$$\mathbf{f}_r{}^T\mathbf{L}^{(1)}\mathbf{f}_r = \frac{1}{2}\sum_{k=1}^{p}\frac{1}{n_k}\sum_{i,j\in k}(x_{ri}-x_{rj})^2 \tag{11}$$

$$\mathbf{f}_r{}^T\mathbf{L}^{(2)}\mathbf{f}_r = -\frac{1}{2}\sum_{k=1}^{p}\frac{1}{n_k(n-n_k)}\sum_{i\in k,j\notin k}(x_{ri}-x_{rj})^2 \tag{12}$$

Combining Eq.(4) and Eq.(11), there is

$$\mathbf{f}_\mathbf{r}{}^\mathbf{T}\mathbf{L}^{(1)}\mathbf{f}_r = \frac{1}{2}\sum_{k=1}^{p}\frac{h_k}{n_k(n-n_k)}\sum_{i\in k,j\notin k}(x_{ri}-x_{rj})^2 \tag{13}$$

Comparing Eq.(12) with Eq.(13), it can be obtained

$$\min(h_k) < -\frac{\mathbf{f}_\mathbf{r}{}^\mathbf{T}\mathbf{L}^{(1)}\mathbf{f}_\mathbf{r}}{\mathbf{f}_\mathbf{r}{}^\mathbf{T}\mathbf{L}^{(2)}\mathbf{f}_r} = \frac{\mathbf{L}_r^{(1)}}{\mathbf{L}_r^{(2)}} < \max(h_k) = h \tag{14}$$

Therefore, from Eq.(14), instead of the feature consistency index in Definition 1, the ratio of two Laplacian scores can also be considered as equivalent estimation of feature consistency. They are over the data graph with the configuration of $\mathbf{W}^{(2)}$ and $\mathbf{W}^{(1)}$. If the feature is consistent with these data graphs, term of $\mathbf{L}_r^{(1)}$ should be smaller and $\mathbf{L}_r^{(2)}$ be larger.

Therefore, from graph theory perspective, the supervised feature selection criterion by Laplacian score can be defined as follows

$$m = -\frac{\mathbf{f}_\mathbf{r}{}^\mathbf{T}\mathbf{L}^{(1)}\mathbf{f}_\mathbf{r}}{\mathbf{f}_\mathbf{r}{}^\mathbf{T}\mathbf{L}^{(2)}\mathbf{f}_r} = \frac{\mathbf{L}_r^{(1)}}{\mathbf{L}_r^{(2)}} \tag{15}$$

Based on the criterion, the feature can be ranked, and a simple searching engine can be defined to select appropriate number of features from the list.

3. Spectral kernel transductive support vector machine

3.1 Density-adjustable spectral clustering

Commonly, the weight of the edge in a Graph is defined by the Euclid distance between the two nodes, and it works very well with the linear data.

But for nonlinear data, such as two clusters shown in Fig.2, data points *a* and *c* belong to the same cluster, and the Euclid distance between points *a* and *b* is less than that between points *a* and *c*. Therefore, it is necessary to measure the similarity of data points in a different way, which can zoom out the path length of those passing through low density area, and zoom in those not. Then the minimum path can be obtained to replace the Euclid distance. It is very useful for machine failure prognosis, because there always exists nonlinear when machine anomaly occurring. Chapelle et.al (2005) proposed a density-sensitive distance based on a density-adjustable path length definition as follows,

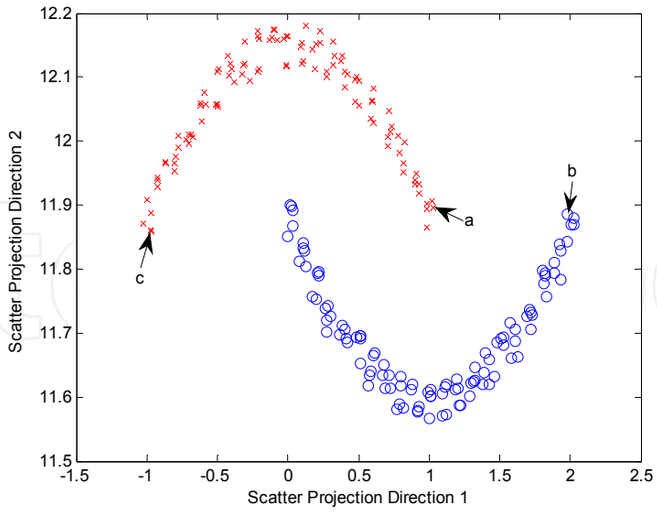


Fig. 2. Scatter Plot of two clusters based on Density-adjustable spectral clustering

$$l(x_i, x_j) = \rho^{dist(x_i, x_j)} - 1 \tag{16}$$

Where $dist(x_i, x_j)$ is the Euclid distance between data x_i and data x_j, and ρ is the density adjustble factor($\rho > 1$). This definition is satisfied with the cluster assumption, and can be used to describe the consistency of data structure by adjusting the factor ρ to zoom out or in the length between the two data points. Therefore, the similarity of the data point x_i and x_j can be expressed as following,

$$s_0(x_i, x_j) = \frac{1}{dsp(l(x_i, x_j)) + 1} \tag{17}$$

Where $dsp(l(x_i, x_j))$ is denoted as the minimum distance between data x_i and x_j, which is the shortest path based on density adjustment.

3.2 Transductive support vector machine

Support vector machine is one of supervised learning methods based on statistical learning theory (Vapnik, 1998). Instead of Empirical Risk Minimization (ERM), Structural Risk Minimization (SRM) is an inductive principle for model selection used for learning from finite training data sets, which enhances the generalization ability of the SVM. The key to SVM is the "kernel tricks", by which the nonlinear map can be realized from low dimensional space to high dimensional space. Therefore, the nonlinear classification task in low dimensional space can be converted to a linear classification, which can be solved by finding a best hyperplane in the high dimensional space.

Considering of 2-class data points, there are many hyperplanes that might classify the data. The best hyperplane is the one that represents the largest margin between the two classes, and the distance from this hyperplane to the nearest data point on each side is maximized.

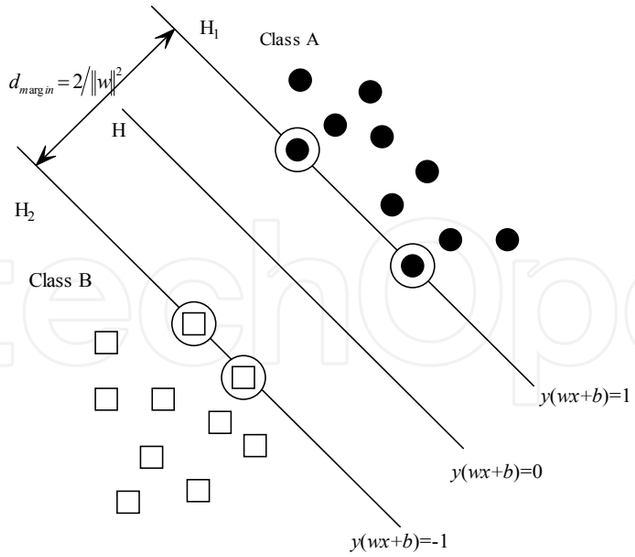


Fig. 3. The Linear Hyperplane of Support Vector Machine

As shown in Fig.3, the data points of Class A are denoted as '•', the others of Class B as '□', and the data points circled by '○' represented support vectors. These data x in the input space are separated by the best hyperplane H

$$\mathrm{y}(\mathbf{w} \cdot x + b) = 0 \tag{18}$$

with the maximal geometric margin

$$\varphi(w) = 2/\|w\|^2 \tag{19}$$

here '$\cdot$' denotes the dot product and **w** is normal vector to the hyperplane, and b is offset from the hyperplane to the margin.

The plane H1 and H2 are also the hyperplanes where the nearest data points to 'H' are located. H1 can be expressed as $\mathrm{y}(\mathbf{w} \cdot x + b) = 1$ and H2 $\mathrm{y}(\mathbf{w} \cdot x + b) = -1$ respectively. It reveals that finding the best hyperplane means minimizing the $\|w\|^2/2$. There are three widely used kernel function as following,

Polynomial Kernel: $K(x, y) = [\langle x, y \rangle + 1]^d$,

Gaussian Kernel: $K(x, y) = \exp(-\| x - y \|^2 / \sigma^2)$,

Hyperbolic: $K(x, y) = \tanh[v(x, y) + c]$.

As for Transductive Support Vector Machine (TSVM), it is one of semi-supervised learning methods, which can combine the labelled data with amounts of unlabelled data co-training. TSVM uses an idea of maximizing separation between labelled and unlabelled data (Vapnik, 1998). It solves

$$\min : \frac{1}{2}\|w\|^2 + C\sum_{i=0}^{l}\xi_i + C^*\sum_{j=0}^{k}\xi_j{}^* \tag{20}$$

$$s.t.: \forall_{i=1}^{l} : y_i(w \cdot x_i + b) \geq 1 - \xi_i \,,\ \forall_{j=1}^{k} : y_j(w \cdot x_j + b) \geq 1 - \xi_j^*$$

$$\forall_{i=1}^{l} : \xi_i > 0 \,,\ \forall_{j=1}^{k} : \xi_j^* > 0$$

Where C and C^* are the penalty factors corresponding to labeled and unlabeled data, ξ_i and ξ_j^* are the slack factors respectively, l is the number of labeled data and k that of unlabeled. These parameters are set by user, and they allow trading off margin size against misclassifying training samples or excluding test samples.

3.3 Density-adjustable spectral kernel based TSVM

Combine the ideas of density-adjustable spectral clustering (Chapelle & Zien,2005) and TSVM, we can get the density-adjustable spectral kernel based TSVM algorithm, called DSTSVM. The data is pre-processed by density-adjustable spectral decomposition, and the processed data is input into the TSVM which is trained by gradient descent on a Gaussian kernel, then the data is classified. The implementation of the DSTSVM algorithm is as following,

Input: n-dimension data X{X_1,…,X_m} (some labelled and others unlabelled)

Parameter: density-adjustable factor ρ, penalty factor C and kernel width σ of the Gaussian kernel. (Set by user)

Output: The label of unlabelled data and the correctness of classification

Step.1 Calculate the Euclid distance matrix S of data X

Step.2 Calculate the shortest path matrix S_0 according to the Eq.16

Step.3 Construct the Graph G based on data matrix S_0. Define the similarity of between nodes as $w_{ij} = e^{-s_0(x_i,x_j)/2\sigma^2}$, and then the degree diagonal matrix can be denoted as $D(i,i) = \sum w_{ij}$.

Step.4 Calculate the Laplacian matrix $L=D^{(-1/2)}WD^{(-1/2)}$ solve the Eigen-decomposition and rearrange the eigenvalue $\{\lambda_1,\ldots,\lambda_n\}$ and corresponding eigenvector {U_1,...,U_n} in descent order.

Step.5 Select the first r nonnegative eigenvectors according to $\left(\sum_{i=1}^{r}\lambda_i \Big/ \sum_{j=1}^{n}\lambda_j\right) \geq 85\%$.

Step.6 Get the new data set as $Y=U_r\Lambda_r^{1/2}$ $\{y_1,\ldots,y_m\}$

Step.7 Train the TSVM by gradient descent using the newdata and then get the classification result.

4. Case study

To demonstrate that the proposed feature selection method and DSTSVM classifier are effective in machine failure prognosis, we applied the methods in feed axis faults feature selection and classification.

4.1 Experiments

Feed axis is one of critical components in a high-precision numerical control machine tool, which always working in conditions such as high speed, heavy duty and large travel distance. This would augment the degradation of mechanical parts such as bearings, ball nuts and so on. From a preventive maintenance perspective, autonomous fault detection and feed axis health assessment could reduce the possibility of causing more severe damage and downtime to machine tool.

TechSolve Inc. collaborated with the NSF Intelligent Maintenance System Center (IMS) to investigate intelligent maintenance techniques for autonomous feed axis failure diagnosis and health assessment. For the investigation, designed experiments were conducted on a feed axis test-bed built by TechSolve. Multiple seeded failures were tested on the system such as axis front and back ball nut misalignment, bearing misalignment and so on (Siegal et.al, 2011). 13 channels (bearing and ball nut accelerometers, temperature and speed; motor power; encode position and so on) data were collected from the test-bed over a period of approximate 6 months. Since all the tests were designed to carry certain failures under different working conditions, the collected information was labeled in terms of the four condition indices including the test index, the load, ball nuts condition, and bearing condition.

Mode	Test index	Table Load	BallNut Misalignment	Bearing Misalignment	Time
1 (Health)	1	0	0	0	2010-10-20
	2	300	0	0	2010-10-22
2 (Failure1)	3	300	0	0.007	2010-11-09
	4	0	0	0.007	2010-11-22
3 (Failure2)	5	0	0.007	0	2010-11-29
	6	300	0.007	0	2010-12-01
4 (Failure1 and Failure2)	7	300	0.007	0.007	2010-12-02
	8	0	0.007	0.007	2010-12-03

Table 1. Eight working conditions of Four modes (Health, Bearing misalignment, Ballnut misalignment, and Combination)

4.2 Feature selection and fault classification

The samples were collected under 4 modes, which were Health, Failure 1(Bearing misalignment $0.007\mu m$), Failure 2 (Ballnut misalignment $0.007\mu m$), and Mode 4 (Failure 1

accompanied with Failure 2). Every mode had two working conditions with load at 0 and 300Kw, and 25 samples at every condition. As for each sample, there were 154 features which contain 117 vibration features (RMS, kurtosis, crest factor at different time periods, and average energy of selected frequency bands) and 37 other features (torque, temperature, position error, and power at different time periods). Therefore, there were totally 200 154-D samples used for investigation.

All the features were evaluated and ranked by Laplacian score using the proposed feature selection criterion. Among 154 features, there were 22 features selected which can reflect the data structure well with the best classification performance, which was shown in Fig.4.

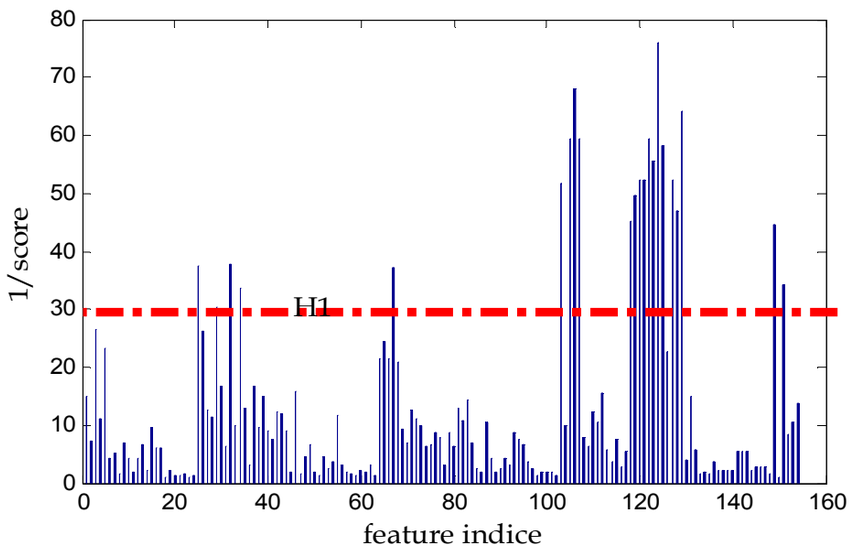


Fig. 4. Features selection based on Laplacian scores

Therefore, the input data dimension can be reduced from 154-D to 22-D. Selecting 25 labelled samples randomly from those 50 22-D samples within every class (totally 100 samples), and the remained 100 samples were regarded as unlabelled ones. Then all these labelled and unlabelled samples were input into the DSTSVM classifier for co-training. This process was repeated for 10 times, and then through 5-fold cross validation, we predicted that which class should the unlabelled samples belong to.

For testing the performance of designed DSTSVM classifier, we reduced the labelled samples to 20 and 10 respectively, and then repeated the procedure above. To verify the effectiveness and correctness, the result was compared with those using SVM (supervised) and TSVM (semi-supervised).

The 10th classification results using the data (10 labelled samples VS 40 unlabelled each class) were shown in Fig.5, Fig. 6, and Fig. 7 respectively.

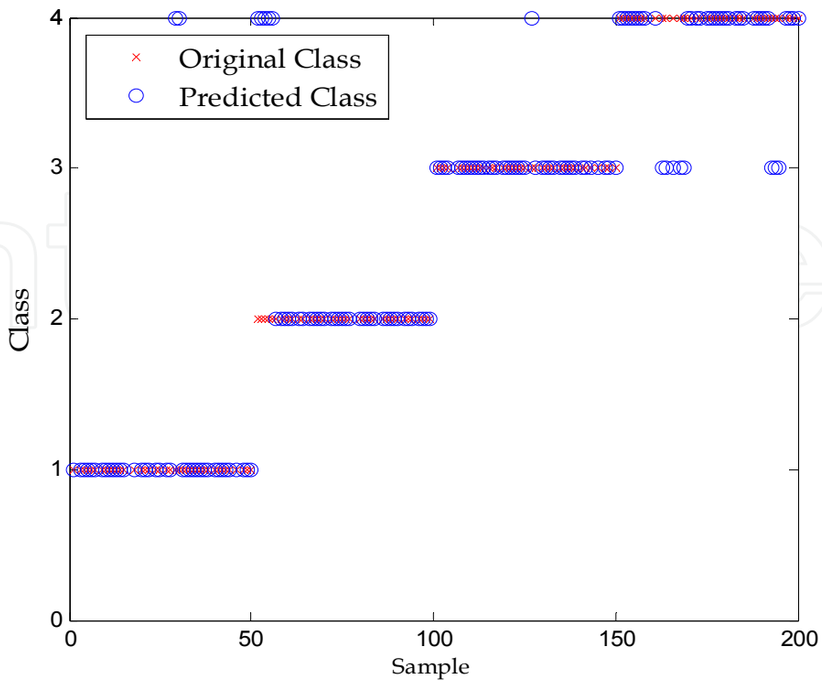


Fig. 5. The Learning result of DSTSVM

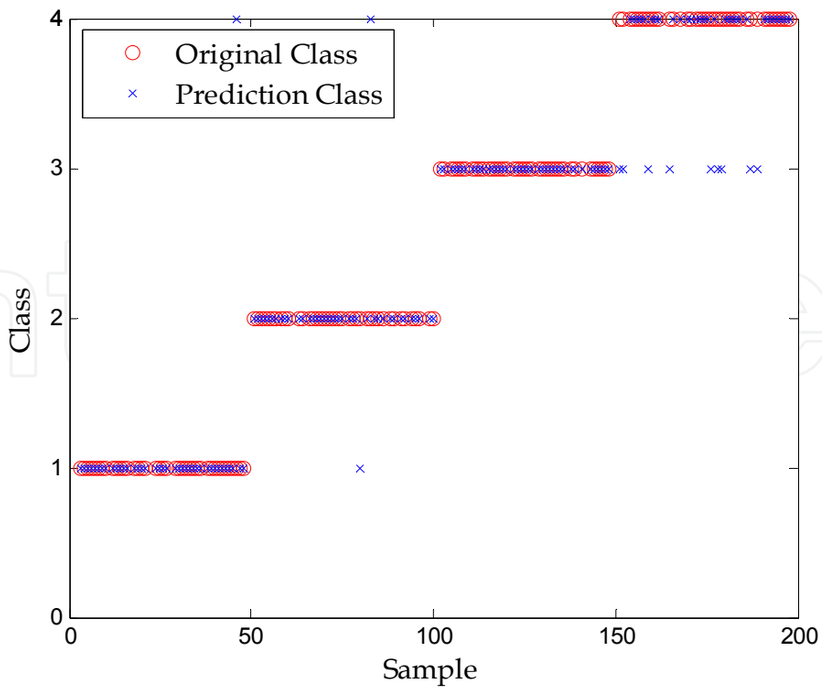


Fig. 6. The Learning result of SVM

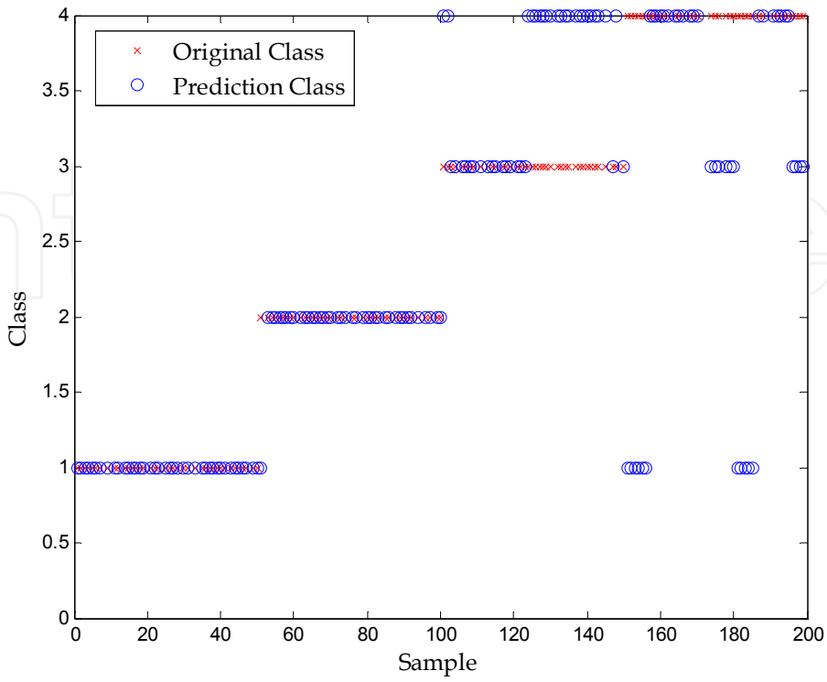


Fig. 7. The Learning result of TSVM

All of the classifiers were trained on Gaussian kernel, and the kernel width σ was set as the optimal value corresponding to the different classifiers. There are two parameters σ and ρ influencing the DSTSVM classification, and the density adjustable factor ρ reflects the data similarity measure, which also affects the Kernel function. In terms of the classification correctness, we can choose the optimal group of these two parameters (ρ, σ). The comparison results under different labelled samples were listed in Table.2.

	DSTSVM		SVM		TSVM	
Labelled vs Unlabelled	Kernel width & density factor (σ,ρ)	Ave Correctness (%)	Kernel width σ	Ave Correctness (%)	Kernel width σ	Ave Correctness (%)
(25vs25)*4	(0.75,2)	91.90	0.5	91.80	0.55	82.50
(20vs30)*4	(0.75,2)	90.92	0.5	90.42	0.55	83.33
(10vs40)*4	(0.75,2)	90.25	0.5	88.88	0.55	85.63

Table 2. The parameters and the average correctness of three classifiers

In Table.2, the average correctness means the average of 10 testing process by 5-fold CV. It can be observed that the proposed method outperforms the TSVM and equals to the supervised SVM under different labelled samples. Moreover, when the labelled data was reduced to 10 samples, it performed better than SVM, which was very meaningful to practical machine failure prognosis applications.

5. Conclusion

The proposed feature selection method can capture the structures of the input data, reduce the dimension of the data and expedite the computation process. More importantly, the classification result is also improved by this feature selection method. Compared with traditional supervised SVM learning and the TSVM semi-supervised learning method, the proposed DSTSVM performed better. Experiment results demonstrate that the proposed DSTSVM method is effective and capable of classifying incipient failures. It has great potential for machine fault prognosis in practice. Based on the current work, the proposed approach can be used to quantify and assure the sufficiency of the data for prognostics applications.

In total, the spectral clustering based method was proposed to evaluate data and to select sensitive features for prognostics, furthermore the spectral kernel based TSVM classifier was also proved to be effective in PHM applications.

6. Acknowledgment

The work described in this paper is supported in part by the National Natural Science Foundation of China (51075150), the Fundamental Research Funds for the Central Universities (2009ZM0091), and Open Research Foundation of State Key Laboratory of Digital Manufacturing Equipment and Technology (DMETKF2009010). The authors would also like to thank Intelligent Maintenance System center in University of Cincinnati and its industry collabrator TechSolve Inc. for the investigation of feed axes system.

7. References

B. Schölkopf, A. Smola, K.R.Muller.(1998). Nonlinear Component Analysis as a Kernel Eigenvalue Problem, *Neural Computation*, Vol.10, No.5,(July 1998), pp.1299–1319, ISSN 0899-7667

Campbell, C. and Bennett, K. (2001). A linear programming approach to novelty detection, *Advances in Neural Information Processing System*, Vol. 14, pp.395-401, ISBN-10 0-262-04208-8, Vancouver, British Columbia, Canada. December 3-8, 2001

Chapelle O, Zien A.(2005). Semi-supervised classification by low density separation, *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics,* pp57-64, ISBN 0-9727358-1-X, Barbados, January 6-8, 2005

D. Dong and T. J. McAvoy.(1996). Batch tracking via nonlinear principal component analysis, *AIChE Journal*, vol. 42,No.8, (August 1996). pp. 2199-2208, ISSN 1547-5905

Harkat, M-F., Djelel, S., Doghmane, N. and Benouaret, M. (2007). Sensor fault detection, isolation and reconstruction using non-linear principal component analysis, *International Journal of Automation and Computing*, Vol. 4, No. 2, (April 2007),pp.149-155, ISSN 1751-8520

I. Jolliffe.(2002)Principal component analysis, *Encyclopedia of Statistics in Behavioral Science*, ISBN 0-387-95442-2. Springer, New York

J. B. Tenenbaum, V.d.Silva and J.C. Langford.(2000). A global geometric framework for nonlinear dimensionality reduction, *Science*, Vol. 290, No. 5500, (December 2000),pp. 2319-2323, ISSN 0036-8075

T. F. Cox. and M. A. A. Cox (2001). Multidimensional scaling, Chapman and Hall(2nd Edition), ISBN 1-584-88094-3, Florida, USA

M. Belkin and P. Niyogi. (2003). Laplacian Eigenmaps for Dimensionality Reduction and Data Representation, *Neural Computation*,Vol.15,No.6 (June 2003), pp.1373-1396, ISSN 0899-7667

M. Ge, R. Du, G. Zhang, Y. Xu. (2004) Fault diagnosis using support vector machine with an application in sheet metal stamping operations. *Mechanical System and Signal Processing*, Vol.18, No.1, (January 2004), pp.143-159, ISSN 0888-3270

M. H. Kaspar and W. H. Ray.(1992). Chemometric methods for process monitoring and high performance controller design, *AIChE Journal*, Vol. 38, No.10, (October 1992), pp. 1593-1608,ISSN 1547-5905

M. Li, J. Xu, J. Yang, D. Yang and D. Wang. (2009). Multiple manifolds analysis and its application to fault diagnosis, *Mechanical systems and signal processing*, Vol.23, No.9, (November 2009) ,pp. 2500-2509, ISSN 0888-3270

P. Nomikos and J. F. MacGregor.(1994). Monitoring batch processes using multiway principal component analysis, *AIChE Journal*, Vol.40, No.8,(August 1994), pp. 1361-1375, ISSN 1547-5905

Q. Jiang, M.Jia, J. Hu and F. Xu. (2009). Machinery fault diagnosis using supervised manifold learning," *Mechanical systems and signal processing*, Vol.23, No.7, (October 2009), pp. 2301-2311, ISSN 0888-3270

R. Kannan, S. Vempala, and A. Vetta.(2004). On clusterings: Good, bad and spectral. Journal of the ACM, Vol.51, No.3, (May 2004), pp.497–515, ISSN 0004-5411

Skirtich, T., Siegel, D. and Lee, J. (2011). A Systematic Health Monitoring and Fault Identification Methodology for Machine Tool Feed Axis. *MFPT Applied Systems Health Management Conference*, pp.487-506, May 10-12, 2011, Virginia Beach, VA, USA

S. T. Roweis and L. K. Saul.(2000). Nonlinear dimensionality reduction by locally linear embedding, *Science*, Vol. 290, No.5500, (December 2000),pp. 2323-2326, ISSN 0036-8075

U. Von Luxburg. (2007). A tutorial on spectral clustering, *Statistics and Computing*, Vol.17, No.4, (December 2007), pp. 395-416, ISSN 0960-3174

U. von Luxburg, M. Belkin, and O. Bousquet.(2008) Consistency of spectral clustering. *Annals of Statistics*, Vol.36, No.2, (April 2008),pp.555–586, ISSN 0090-5364

Vapnik, V. (1998). Statistical Learning Theory. Wiley, ISBN 0-471-03003-1, New York, USA

W.C Sang, C Lee, J Lee, H Jin, I Lee. (2005). Fault detection and identification of nonlinear processes based on kernel PCA, *Chemometrics and intelligent laboratory systems*, Vol.75, No.1, (Januray 2005), pp.55-67, ISSN 0169-7439

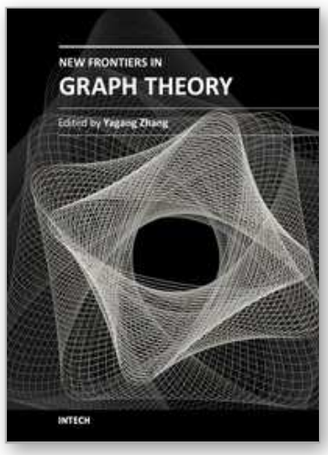
W. Yan and K. F. Goebel. (2005). Feature selection for partial discharge diagnosis, *Proceedings of the 12th SPIE: Health Monitoring and Smart Nondestructive Evaluation of Structural and Biological Systems IV*, Vol.5768, pp. 166-175, ISBN 0-8194-5749-3, March 19 - 22, 2007, San Diego , California,USA

X. He, D. Cai, P. Niyogi.(2006). Laplacian score for feature selection, *Advances in neural information processing systems*, Vol.18, p507-515, ISBN 0-262-19568-2, December 4-7,2006, Vancouver, British Columbia, Canada

Y Lei, M Zuo.(2009). Gear crack level identification based on weighted K nearest neighbor classification algorithm, *Mechanical Systems and Signal Processing,* Vol.23,No.5, (July 2009),pp.1535-1547, ISSN 0888-3270

Ypma, A. (2001). Learning methods for machine vibration analysis and health monitoring, *PhD Dissertation*, ISBN 90-9015310-1, Delft University of Technology, Delft, Netherlands.

Z. Zhao and H. Liu, (2007). Spectral feature selection for supervised and unsupervised learning, Proceedings of the 24th International Conference on Machine Learning , pp.1151-1157. ISBN 978-1-59593-793-3, June20-24 ,2007, Corvallis, OR, USA



New Frontiers in Graph Theory

Edited by Dr. Yagang Zhang

ISBN 978-953-51-0115-4
Hard cover, 526 pages
Publisher InTech
Published online 02, March, 2012
Published in print edition March, 2012

Nowadays, graph theory is an important analysis tool in mathematics and computer science. Because of the inherent simplicity of graph theory, it can be used to model many different physical and abstract systems such as transportation and communication networks, models for business administration, political science, and psychology and so on. The purpose of this book is not only to present the latest state and development tendencies of graph theory, but to bring the reader far enough along the way to enable him to embark on the research problems of his own. Taking into account the large amount of knowledge about graph theory and practice presented in the book, it has two major parts: theoretical researches and applications. The book is also intended for both graduate and postgraduate students in fields such as mathematics, computer science, system sciences, biology, engineering, cybernetics, and social sciences, and as a reference for software professionals and practitioners.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen