The Design of Bearing Fault Diagnosis Classifier Based on the Bayesian Classification

Chunhe Shi[1, a], Xiaowei Han[1, b], and Zhen Li[2, c]

[1]School of Information Engineering, Shenyang University, Shenyang 110044, China

[2]Northeast Regional Air Traffic Management Bureau of CAAC, Shenyang 110043, China

[a]schsydx@163.com, [b]hxw69@163.com, [c]zlalbert@126.com

Keywords: PCA, Bayesian, Fault Detection

Abstract: The advantages of Principal Component Analysis (PCA) in dimension reduction are obvious, while the classifying methods are various. PCA, applied to Bayesian classification with the minimium risk, is adopted to realize classifying and testing the learning samples in this paper, eventually accurately realizing the classifier training towards the new input sample and then detecting the testing effect.

1. Introduction

In recent years, the Bayesian has become a primary tool of data mining and knowledge discovery, and achieved good applying effect in classification, clustering, forecasting, and derivation of the rules, etc. Therefore, this paper uses Bayesian classifier to achieve the rolling bearing fault diagnosis. PCA (Principal Component Analysis) is a technique of converting the data set to minor dimensions and components containing all or major information of the original data set via a reversible linear transformation. By using PCA, the complex data will be simplified, thence, it is widely used in data mining, pattern recognition, signal evaluation, signal detection, image coding, and other fields. PCA's principles are as follow:

Let x be a random vector representing m-dimension. Assuming that the x mean is zero, i.e.:

$$E(x) = 0 \tag{1}$$

Let w be m-dimensional unit vector, and x the mapped in ω. The mapping is defined as the inner product of the vector x and ω, expressed as:

$$y = \sum_{k=1}^{n} W_k X_K = W^T X \tag{2}$$

Satisfying the constraints

$$\|w\| = (w^T w)^{1/2} = 1 \tag{3}$$

The purpose of PCA is searching for a weight vector w, which maximizes the value of the expression:

$$E[y^2] = E[(w^T x)^2] = w^T E[xx^T]w = w^T C_x w \tag{4}$$

$$C_x W_j = \lambda_j w_j \qquad j = 1,\ 2\cdot,\ m, \tag{5}$$

i.e., w which maximizes the value of equation (4) is the corresponding eigenvector to the largest eigenvalue of the matrix.

Considering the advantages of PCA, it is herein introduced to reduce and demensions of data sets and process the data, in order to improve the efficiency of Bayesian classifier and simplify the structure.

2. Basic Principles of PCA and Bayesian Classification

2.1 Basic Principles of PCA

PCA is a statistical analysis method on which multiple features are mapped as a few comprehensive characteristics. Adopting dimension reduction, PCA finds several variables to represent the multiple original features, making these variables reflect the original information to the utmost and irrelavant to each other, so as to achieve the objective of simplification.

Its basic idea is to attempt to reassemble the multiple related indexes $X_1,X_2,...,X_P$ (for example p indexs) into a reduced number of disparate set of comprehensive index F_m to replace the original ones. How to extract Indexes, so that it can both reflect to the maximum the information represented by the variable X_p, and ensure new indexes' irrelavance (without information overlapping).

Suppose that F_1 represents the principal component index formed by the first linear combination of the original variable, i.e. $F_1 = a_{11}X_1 + a_{21}X_2 + ... + a_{p1}X_p$, the amount of information extracted by each principal component can be measured via its variance. The higher the variance Var (F_1) is, the more information F_1 contains. Usually, the first principal component F_1 is suppoed to be the most informative, therefore, the variance of F_1, selected from all the linear combinations, ought to be the largest among $X_1,X_2,...,X_P$, thus F_1 is named the first principal component. If the first principal component is not able to represent the original information of p indexes, consider selecting a second principal component index F_2. To effectively reflect the original information, the information contained in F_1 does not necessarily appear in F_2, i.e., F_2 and F_1 keep independent and irrelavant. Expressed in mathematics language, its covariance Cov (F_1, F_2) = 0, so F_2 is of the maximium variance among all the linear combination, say, X_1, $X_2,...,X_P$, thus F_2 is named the second principal component, and by analogy F_1, F_2, ..., F_m are the first, second, ..., m-th principal components of the original variables X_1, X_2, ... X_P.

According to the above analysis,

(1) the F_i and F_j are irrelavant, namely Cov (F_i, F_j) = 0.

(2) F_1 is of the maximium variance among all the linear combinations (the coefficients satisfy the above requirements), say, F_1 are X_1, X_2, ..., X_P; i.e., Fm is of the maximium variance among all the linear combinations X_1, X_2, ..., X_P irrelavant to $F_1,F_2,.....,F_{m-1}$. F_1, F_2, ..., F_m ($m \leq p$) are the new constructed variables, i.e. the first, second, ..., m-th principal components of the original variables.

2.2 Basic Principles of Bayesian Classification

Due to the complexity of objective things, the risks of classifier judgement are not identical. The risk of misjudging cancer cells as normal ones, for example, is greater than that of miscalculating normal cells as cancer ones. Therefore, the concept of risk is introduced in Bayesian classifiers. In practical application, the magnitude of risks is identified according to specific situations, usually represented by a set of coefficients C_{ij}. C_{ij} refers to the risk when classifier is classified as ω_i by the recognition samples, while the actual classificaton of the sample is ω_j. The basic idea of designing a classifier with minimal risk is to calculate the conditional risk of classifying x into ω_i via the posterior probability calculation.

Comparing each value of R_i (x), the classification corresponding to the minimum is the result. Since the standard of evaluating classifiers is average risk, this classifier is of the minimum average risk. In practical application, posterior probabilities are difficult to obtain. According to the number of pattern classificatoin and value-reading pattern of C_{ij} categories, a variety of classifiers are able to be designed, for example, when patterns contain two classifications, discriminant functions are

$$g_1(x) = (C_{21} - C_{11})P(x|w_1)P(w_1)$$
$$g_2(x) = (C_{12} - C_{22})P(x|w_2)P(w_2) \quad (6)$$

Select C_{11}, C_{22} as zero, C_{12} and C_{21} as 1, and this brings about a classifier of two classifications with minimum error probability. Actually, the minimum error probability classifier is a special case of minimal risk classifiers.

3. Experimental Procedure

(1) Selecting all samples of different classifications to form Matrix $X_{n\times N}$, with the testing sample is $X_{n\times 1}$.

(2) Calculating the covariance matrix $S_{n\times n}$ of $X_{n\times N}$.

(3) Calculating the eigenvalue $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ of $S_{n\times n}$.

(4) According to a certain contribution rate, selecting the first m columns of $C_{n\times n}$, constituting $C_{n\times m}$.

(5) Calculating the sample component $X_{m\times n}= CT_{n\times m}X_{n\times N}$ from sample gallery and the principal component $X_{m\times 1}= CT_{n\times m}X_{n\times 1}$.

(6) Calculating the mean value of each classification of digital samples

$$\overline{X^{(w_i)}} = \frac{1}{N_i}\sum_{X\in w_i} X = (\overline{x_1^{(w_i)}}, \overline{x_2^{(w_i)}}, \cdots, \overline{x_n^{(w_i)}})^T \qquad i = 0,$$

In the above equation, Ni represents the number of samples of the w_i classification and n represents the number of features.

(7) Evaluating covariance matrix of each classification.

$$S_{jk}^i = \frac{1}{N_i - 1}\sum_{l=1}^{N_i}(x_{lj} - \overline{x_j^{(w_i)}})(x_{lk} - \overline{x_k^{(w_i)}}) \qquad j,\ k = 1, 2\cdot,\ \imath$$

In the above equation, l stands for the serial number of samples in the w_i classification, in which $l = 0,1,2,\cdots,N_i$

x_{lj} represents the j-th feature value of the l-th sample in w_i classification;

$\overline{x_j^{(w_i)}}$ represents the average value of the j-th feature of a N_i number of samples in w_i classification;

x_{lk} represents the k-th feature value of the l-th sample in w_i classification;

$\overline{x_k^{(w_i)}}$ represents the average value of the k-th feature of a N_i number of samples in w_i classification;

x_{lj} represents the j-th feature value of the l-th sample in w_i classification;

(8) Calculating the inverse matrix S_i^{-1} of each classification's covariance matrix, and the determinant $|S_i|$ of the covariance matrix.

(9) Evaluating the prior probability of each classification.

$P(w_i) \approx N_i / N \qquad i = 0, 1$

In the above equation, P(w_i) is the prior probability of number classification; N_i is the number of samples of number i; and N is the total number of samples.

(10) Calculating the posterior probability $P[i]$, $i = 0,1.$

$$P[i] = -\frac{1}{2}(X - \overline{X^{(w_i)}})^T S_i^{-1}(X - \overline{X^{(w_i)}}) + \ln P(w_i) - \frac{1}{2}\ln|S_i|$$

(11) Defining the array of loss as loss[2][2], and setting the initial value as

$$loss[i][j] = \begin{cases} 0, i = j \\ 1, i \neq j \end{cases}$$

(12) Calculating the loss of each classification

$$risk[i] = \sum_{j=0}^{1} loss[i][j]p[j]$$

(13) Finding the classification corresponding to the minimum loss, accordingly identifying this classification as the one that testing samples belong to.

4. Experimental Results and Analysis

Using PCA after the dimension structure of the Bayesian classification and unused dimension data learning results on classification accuracy differ in structural classification accuracy is much higher than other classifiers, also using the dimension data have nodes small, simple structure and high efficiency. Use the test sample test result, but through this classifier classification classification is fully consistent with the test sample, this classifier is very accurate.

5. Conclusion

Bayesian classifier of low efficiency of shortcomings, presented data set through the PCA principal components analysis dimensionality reduction. Reduce data, then data dimension reduction method for Bayesian learning, improved operational efficiency and by increasing the efficiency of correctness to ensure better results. In addition, this method is simple, high recognition rate, in real time, trained networks have very good classification abilities.

References

[1]. Yang Shuying. Pattern recognition and intelligent-Matlab technologies. Electronic Industry Press, 2008, p. 132-140.

[2]. Zeng Huanglin, Yu Jue-bang, Zeng Qian. Feature simplification based on principal components analysis. Journal of Sichuan Institute of light industry, Vol. 12(1999) No. 1, p. 1-4.

[3]. N.Huang. Based on PCA Bayesian network classifiers. Mathematical statistics and management, Vol. 18(1999) No.5, p. 44-52.

[4]. Chen Y D, Du R. Fault features of large rotating machinery and diagnosiusing sensor fusion. Journal of Sound and Vibration, Vol. 188(1995) No.2, p. 227-242.