Study of feature vector discriminability optimization for classification based on PCA and MDA

Jiang Xiangdong[1,2,a], Tang Jiansheng[2,b], Xiao jigang[2], Yunji Jin[2], Zou Jinshun[2]

[1]Harbin Engineering University, Heilongjiang Harbin, China

[2]Science and Technology on Underwater Acoustic Antagonizing Laboratory, Beijing, China

[a]Jiangxd0618@163.com, [b]Tangjs.npu @gmail.com

Abstract. To solve the problem of weak discriminability of the feature vector for underwater acoustic classification, a method of feature differentiation optimization based on PCA and MDA analysis was proposed in this paper. It can enhance the differentiation by optimal mapping the feature vectors to transform space. The data processing results proved the method is feasible and has the advantage of feature dimension reducing that is useful in practice.

Introduction

Classification is normally implemented in tow steps: feature extraction and pattern classification. A ideal feature extractor can greatly simplify the work of second step (pattern classification),which can improve the similarity of the feature vector from the same class and enhance the differentiation of the feature vectors from different l classes. So it is critical to choose a discriminable feature parameter or parameter vector. At present, based on the signal form the hydrophone or sensor array, many feature extraction methods had been proposed, such as line-spectrums, auditory feature, wave structure, spectrum image feature, high order statistical feature etc. Some features correspond to the traits of the mechanical system, which are normally useful for recognizing a special target, especially supported by database .Some features are statistical for general types classification, which have to be acquired by analyzing a large amount of samples of different types.

However, due to the complexity of so many kinds of underwater vehicles and the disturbing effect of time varied environment, It is hard to achieve a ideal passive recognition in practice. In fact, feature extraction, classifier design and feature optimization are correlated with each other. Based on the conventional feature extraction method, the problem of how to select optimum feature vector, how to reduce the dimension and ensure the classification performance at the same time, how to fusion different kinds of features and classifiers has become a rising research aspect.

In this paper, to solve the problem of weak differentiation of the feature vector, a method of feature differentiation optimization based on PCA （principal component analysis and MDA(multi-classes discriminant analysis) was proposed. For high dimension feature, we use PCA linear transform to get a set of main feature parameters and ignore the unimportant parts. For multi-classes condition, to enhance the discriminability, we use MDA transform to mapping the feature vectors to transform space. Simulation and data processing result statistic results proved the method is feasible.

Main feature selection of by PCA process

We define x_k as the D dimension feature vector for the k th class targets. For the one certain class targets, we have n feature vectors $\{x_k | x_k \in M_i, k = 1, 2 \cdots n\}$ for n targets. In training the classifier, we always hope to get features which can mostly represent the main characteristics of all targets in that class, and also differentiable and low dimensional for computing. The unimportant features which may cause illegibility should be ignored.

The within-class scatter matrix for k th class targets is:

$$S_i = \sum_{k=1}^{n} (x_k - m)(x_k - m)^T \tag{1}$$

In equation (1), m is the mean value of the feature vector samples of the k th class targets.

$$m = \frac{1}{n} \sum_{k=1}^{n} x_k \tag{2}$$

S_i is a real symmetrical matrix which can be transformed to a triangle matrix:

$$\Lambda = P^T S_i P = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \cdots & \\ & & & \lambda_D \end{bmatrix} \tag{3}$$

P is a matrix which is composed of the character vectors of S_i, $p = \{e_1, e_2 \cdots e_D\}$. Any target's feature vector within k th class can be expressed by linear combination of $e_1, e_2 \cdots e_D$.

$$x_k = m + \sum_{k=1}^{D} a_k e_k \tag{4}$$

To select the main differentiable features, we normally reduce the dimension from D to $d(d < D)$,

$$x_k \approx x'_k = m + \sum_{k=1}^{d} a_k e_k \tag{5}$$

It had been proved that error function J_d in equation (6) bas minimum mean square value.

$$J_d(A,e) = \sum_{k=1}^{n} \left\| \left(m + \sum_{k=1}^{d} a_k e_k \right) - x_k \right\|^2 \to \min \tag{6}$$

For certain target class, the original feature vector of each target in this class forms a D dimensional distribution, the character vectors $e_1, e_2 \cdots e_D$ of S_i represent the main direction of the distribution. PCA transform choose a set of main values in this main direction.

E.g, supposed x_k is D dimensional which is composed of the frequency of the line-spectrums of the target underwater radiated noise. Through PCA process, we get a vector that composed the main line-spectrums which have relative larger energy and higher signal to noise ratio.

Feature discriminability enhance by lda process

For a certain class target, as we discussed in section II, we use PCA method to extract the main feature. In this section, for multi-classes condition, we try to find some kind of transformation to assemble the same class of targets feature and partition different classes of targets feature markedly. So we can get better performance by optimizing the feature vectors discriminability in the training the classifier. E.g in fig.1, for two classes of targets, each target has two dimensional feature vectors, if we map the feature vector points to the direction of Ψ_2 rather than Ψ_1, better performance can be obtained for classification. Mapping transformation greatly simplifies the work of classification.
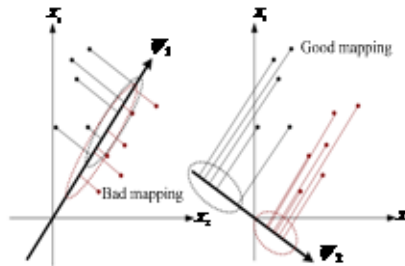

Fig.1 sketch of feature differentiation contrast by mapping to different direction in two dimension

Suppose X is feature vector aggregation for n targets, $X = \{x_k, k = 1,2...N\}$, which can be sorted to M classes. Each class including n_i targets. We resort to MDA analysis to find a optimized

transform matrix W , which can transform the feature vectors X to Y .

$$Y = (y_1, y_2 \cdots y_n) = W^T X = (w_1, w_2, \cdots w_n)^T X \qquad (7)$$

We denote M_i as i th class targets in X , the within-class scatter matrix is the same form as equation (1):

$$S_i = \sum_{x_i \in M_i} (x_i - m_i)(x_i - m_i)^T \qquad (8)$$

For all of the M classes target feature vector samples, the sum of within-class scatter matrix is:

$$S_w = \sum_{i=1}^{M} S_i$$
$$= \sum_{i=1}^{M} \sum_{x_i \in M_i} (x_i - m_i)(x_i - m_i)^T \qquad (9)$$

For all of the M classes target feature vector samples, the sum of between-class scatter matrix is:

$$S_B = \sum_{i=1}^{M} \sum_{x_i \in M_i} (m_i - m)(m_i - m)^T \qquad (10)$$

For all of the M classes target feature vector samples, the total class scatter matrix is:

$$S_T = \sum_{x_i \in M_i} n_i (x_i - m)(x_i - m)^T$$
$$= S_W + S_B \qquad (11)$$

In equation (8),(9),(10)and (11)，m_i is mean of feature vector samples for i th class. m is mean of feature vector samples for all target in X .

$$m_i = \frac{1}{n_i} \sum_{x_i \in M_i} x_i \qquad (12)$$

$$m = \frac{1}{N} \sum_{x_i \in M_i} n_i x_i \qquad (13)$$

In condition of $M = 2$, equation (9),(10) can be simplified:

$$S_W = S_1 + S_2 \qquad (14)$$

$$S_B = (m_1 - m_2)(m_1 - m_2)^T \qquad (15)$$

We has to find a transform matrix W for equation (7) which can maximize the ratio of S_B to S_W ,which is called cost function.

$$J(W) = \frac{W^T S_B W}{W^T S_w W} \qquad (16)$$

It had been proved the character vectors w_i in (17) compose the matrix W .

$$S_B w_i = \lambda_i S_W w_i \qquad (17)$$

First the character value λ_i can be calculated:

$$\left| S_B - \lambda_i S_W \right| = 0 \qquad (18)$$

For each λ_i , the character vector can be calculated:

$$(S_B - \lambda_i S_W) w_i = 0 \qquad (19)$$

In condition of $M = 2$:

$$w_i = S_W^{-1}(m_1 - m_2) \qquad (20)$$

So, we get $W = \{w_1, w_2, \ldots w_n\}$

Process of feature vector discriminability optimization and performance test

Based on II and III, we made up a process flow as shown in fig.2. The performance of the classifier is compared by using original features. Only PCA，PCA and MDA.
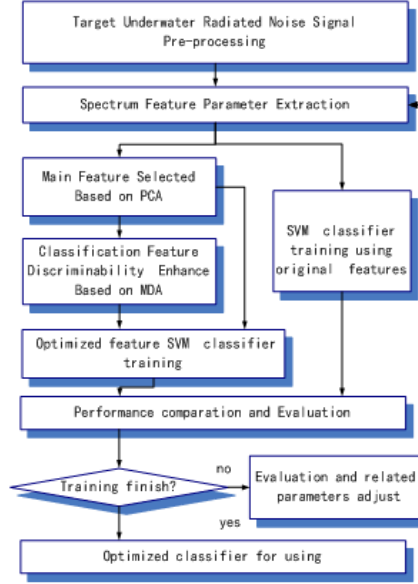
Fig.2 process flow of feature discriminability optimization and performance test

The recorded time domain radiated noise signals of each target are unequal. In the signal of each target are segmented by hanning window function. The time length T of the window is decided by the frequency resolution of power spectrum. In training and test, we use 1second time length window to get 1Hz resolution. Then we average the spectrum in 5 seconds. The averaged spectrums are prepared for feature extraction. For a relative longer signal, we get many averaged spectrums for one target. The amounts of averaged spectrums of each target may be different according to different time length. Some are used for training, others may be used for testing.

In spectrum feature parameter extraction step, at first the power spectrum density (PSD) is calculated based on FFT. Then the following feature parameters are calculated:

(a) X_1: sub-band source level, the band width is ΔF and center frequency is $f_i (1 < i < n1)$

$$X_1 = [SL(f_1), SL(f_2), \cdots SL(f_{n1})]$$

$$SL(f_i) = \int_{f=f_i-\Delta F/2}^{f_i-\Delta F/2} PSD(f) = \int_{f=f_i-\Delta F/2}^{f_i-\Delta F/2} \frac{1}{2\pi} \int_{t=0}^{T} s(t)e^{-2\pi f t} dt \qquad (20)$$

(b) X_2: Number of line spectrums in each sub-band ΔF

$$X_2 = [N_{lines}(f_1), N_{lines}(f_2), \cdots N_{lines}(f_{n1})] \qquad (21)$$

In line spectrum extraction, we mainly consider the amplitude and frequency stability. The line spectrum SNR should be above the extraction threshold and it's frequency is relatively stable. Then we count the line spectrums in each sub-band .

(c) X_3: Energy of line spectrums in each sub-band ΔF

The sum of the amplitude of the line spectrums in each sub-band $E_{lines}(f_i)$

$$X_3 = [E_{lines}(f_1), E_{lines}(f_2), \cdots E_{lines}(f_{n1})] \qquad (22)$$

The dimension $n1$ of X_1, X_2, X_3 is same which decided by the signal band B and ΔF, $n1 = B/\Delta F$. To remove the effect of the signal input amplify and the difference of sensors' sensitivity, we calculate the ratio of $E_{lines}(f_i)$ to total energy.

(d) X_4: Main line spectrum frequency. We choose the line spectrums which SNR are higher than others. E.g, the selected lines with square mark .In the software program we search the n_2 highest SNR lines automatically based on the PSD then we compose X_4 by their frequencies.

$$X_4 = [F_{lines}(f_1), F_{lines}(f_2), \cdots F_{lines}(f_{n2})] \qquad (23)$$

The feature vector for classification is arranged as:

$$X_{1 \times D} = [X_1, X_2, \cdots, X_4], D = 3n_1 + n_2 \qquad (24)$$

In the test , the dimension $n1 = 12, n2 = 9, D = 45$. After we get X for all signal segments, we use 30% of them to train the SVM classifier and the left for testing. The results are:

(1) when only PCA is used, we find the correct probability is higher than using the original feature directly. As shown in fig.3,there is a best dimension ,below which the performance get better when the main discriminable feature is selected and the confusing features is ignored. When the dimension is reduced to a certain number, the performance will fall because some part the main feature is lost.

(2) When both PCA and MDA are used, the classify performance get much better. But near and above the best dimension, the improvement is little. When feature dimension is over reduced, the MDA has no effect because the very low dimension feature vectors lost too much information.
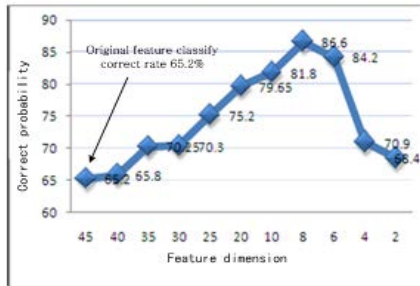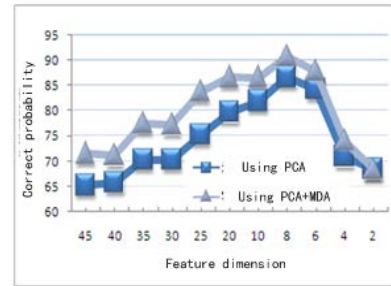

Fig.3 performance comparation (PCA)   Fig.4 performance comparation(PCA+MDA)

Conclusion

In classification, feature discriminability optimization and dimension reducing may be contrary. The method proposed in this paper solves the problem through PCA and MDA process for spectrum features. Foe the targets samples used in this paper, the PCA process ignored unimportant features which may cause illegibility. The classification performance of SVM improved by more than 10% .When a best dimension is selected, the performance improved about 20%. Dimension reducing is very necessary for microprocessors in portable equipment. Further work should be done to test the adaptability for other features.

References

[1] Richard O. Duda, Peter E Hart, David G. Stock, Pattern Classification(Second Edition) ;

[2] Jackson J E. A user's guide to principal components. Hoboken, New Jersey: John Wiley & Sons Inc. , 2003: 1-6.

[3] Ben F, Liu J. Binary tree of SVM : A new fast multiclass training and classification algorithm [J]. IEEE Transactions on Neural Net works, 2006, 17(3) : 696- 704.

[4] Geoffrey J. McLachlan. Discriminant analysis and statistical pattern recognition.Wiley, NewYork, 1992