COMMON VECTOR APPROACH IN MULTI-CLASS PROBLEMS AND ITS RELATION TO FISHER'S LINEAR DISCRIMINANT ANALYSIS

M. Bilginer Gülmezoğlu[1], Vakıf Dzhafarov[2], Rifat Edizkan[1] and Atalay Barkana[1]
bgulmez@ogu.edu.tr, vcaferov@anadolu.edu.tr, redizkan@ogu.edu.tr, abarkana@ogu.edu.tr
[1] *Osmangazi University, Electrical and Electronics Engineering Department, Eskişehir, Turkey*
[2] *Anadolu University, Department of Mathematics, Eskişehir, Turkey*

Keywords: Speech recognition, multi-class problems, common vector approach

ABSTRACT

The idea of the common vector of a class is to find a unique common vector which represents the common properties or invariant features of the class. The previous derivations of the common vector of a class did not consider the distributions of the feature vectors in the other classes. The derivation of the common vector considering the within- and between-class variations in the multi-class case is given in this paper. The Fisher's discriminant analysis is also given and compared with the proposed method. The new idea of the common vector is applied to the isolated word recognition problems and the recognition rates are provided.

I. INTRODUCTION

The common vector derived in the previous papers [1,2] belongs to the within-class or the single class distributions. Since it did not include the between-class or inter-class distributions, one may tend to think that it should produce low recognition rates. But the experimental study showed that very high recognition rates can be obtained with the previous common vector approach (CVA), i.e. 100% for the training set, 94% for the test set of 17 isolated words. However when the within-class and between-class distributions of the classes are similar to the case shown in Fig. 1, the CVA with within-class distributions does not obviously work for the recognition purposes.

Therefore a more general approach considering the between-class distributions as well as the within-class distributions similar to the Fisher's linear discriminant analysis (LDA)[3] is essential for the derivation of the common vector of a certain class. In this paper, it is shown that the common vector derived by considering the within-class and between-class distributions in multi-class case can be effectively used in the isolated word recognition problems.

In Section II, the CVA for the multi-class problems is given. The experimental study with the classification criteria and the recognition rates obtained for the LDA and CVA methods are given in Section III. Section IV gives the conclusion and discusses the results.

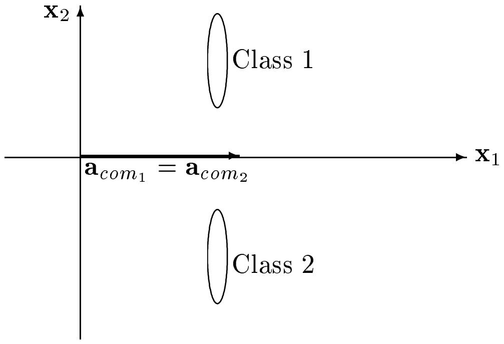

Figure 1. Equal probability density contours of two classes with the same common vectors.

LINEAR DISCRIMINANT ANALYSIS (LDA)

The purpose is to pre-process the data so as to reduce its dimensionality before applying a classification problem. It also establishes a discriminant function. In this section, the generalization of the Fisher's discriminant to several classes is examined [5]. It is assumed that the dimension of the input space is greater than the number of classes, so that *d*>*c*. Totally *d*' discriminant functions can be introduced as

$$y_k = \mathbf{w}_k^T \mathbf{x} \qquad (k = 1, 2, ..., d')$$

where $\mathbf{x}$ is any feature vector and y_k can conveniently be grouped together to form a vector $\mathbf{y}$. The weight vectors $\{\mathbf{w}_k\}$ can be considered to be the rows of a matrix $\mathbf{W}$, so that

$$\mathbf{y} = \mathbf{W}\mathbf{x}$$

The generalization of the within-class covariance matrix to the case of c classes gives

$$\mathbf{S}_W = \sum_{k=1}^{c} \mathbf{\Phi}_k$$

where

$$\boldsymbol{\Phi}_k = \sum_{n\in C_k} (\mathbf{x}^n - \mathbf{m}_k)(\mathbf{x}^n - \mathbf{m}_k)^T$$

and

$$\mathbf{m}_k = \frac{1}{N_k}\sum_{n\in C_k}\mathbf{x}^n$$

where N_k is the number of the feature vectors in class C_k. The between-class covariance matrix is defined as

$$\mathbf{S}_B = \sum_{k=1}^{c} N_k(\mathbf{m}_k-\mathbf{m})(\mathbf{m}_k-\mathbf{m})^T$$

where $\mathbf{m}$ is the mean of the total data set

$$\mathbf{m} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}^n = \frac{1}{N}\sum_{k=1}^{c}N_k\mathbf{m}_k$$

and $N = \sum_k N_k$ is the total number of data points.

Fisher's criterion (metric) can be written in the form of

$$\mathbf{F}_{Fisher} = \frac{\mathbf{W}S_B\mathbf{W}^T}{\mathbf{W}S_W\mathbf{W}^T}. \qquad (1)$$

The maximization of this criterion yields the result that the weight vectors $\{\mathbf{w}_k\}$ are determined by those eigenvectors of $\mathbf{S}_W^{-1}\mathbf{S}_B$ which correspond to the d' largest eigenvalues.

II. COMMON VECTOR FOR MULTI-CLASS CASE

For a two-class case, we should find such an r-dimensional subspace ($r < d$) that the projections of the feature vectors of the classes C_1 and C_2 must be close to their own means respectively. Within the same subspace, the projections of the feature vectors of the first class C_1 must be far away from the mean of the second class C_2, also the projection vectors of the second class C_2 must be far away from the mean of the first class C_1.

If there are only two classes as in the LDA with the means,

$$\mathbf{m}_1 = \frac{1}{N_1}\sum_{n\in C_1}\mathbf{x}^n \qquad \text{and} \qquad \mathbf{m}_2 = \frac{1}{N_2}\sum_{n\in C_2}\mathbf{x}^n$$

then the within-class optimization problem can be written as

$$\mathbf{F}_W{=}\sum_{n\in C_1}\sum_{j=1}^{r}\mathbf{u}_{jW}^T(\mathbf{x}^n{-}\mathbf{m}_1)(\mathbf{x}^n{-}\mathbf{m}_1)^T\mathbf{u}_{jW}{=}\sum_{j=1}^{r}\mathbf{u}_{jW}^T\boldsymbol{\Phi}_1\mathbf{u}_{jW}$$

where $\boldsymbol{\Phi}_1$ is the within-class covariance matrix of the class C_1 and $\mathbf{u}_{jW}$'s are the basis vectors of the subspace. The metric $\mathbf{F}_W$ should be minimized.

The insertion of $\mathbf{u}_i^T\mathbf{u}_j$={1 if i=j ; 0 if i≠j} including the Lagrange multipliers λ_j for each $\mathbf{u}_j$, the minimization problem yields

$$\boldsymbol{\Phi}_1\mathbf{u}_{jW} = \lambda_{jW}\mathbf{u}_{jW}$$

where the Lagrange multiplier λ_{j_W}'s turn out to be the smallest eigenvalues of the covariance matrices $\boldsymbol{\Phi}_1$.

The between-class optimization problem can also be written as

$$\mathbf{F}_B{=}\sum_{n\in C_1}\sum_{j=1}^{r}\mathbf{u}_{jB}^T(\mathbf{x}^n{-}\mathbf{m}_2)(\mathbf{x}^n{-}\mathbf{m}_2)^T\mathbf{u}_{jB}{=}\sum_{j=1}^{r}\mathbf{u}_{jB}^T\boldsymbol{\Phi}_B\mathbf{u}_{jB}$$

where $\boldsymbol{\Phi}_B$ is a different kind of the between-class covariance matrix of the LDA which shows the variance of the feature vectors of the class C_1 to the average feature vector of the class C_2. In this case, the metric $\mathbf{F}_B$ should be maximized.

The maximization problem yields

$$\boldsymbol{\Phi}_B\mathbf{u}_{jB} = \lambda_{jB}\mathbf{u}_{jB}$$

where the Lagrange multiplier λ_{jB}'s turn out to be the largest eigenvalues of the covariance matrices $\boldsymbol{\Phi}_B$.

The within-class and between-class optimization problems can be combined to obtain the following metric:

$$\mathbf{F}_{com}{=}\tfrac{\mathbf{F}_W}{\mathbf{F}_B}{=}\frac{\sum_{j=1}^{r}\mathbf{u}_{jW}^T\boldsymbol{\Phi}_1\mathbf{u}_{jW}}{\sum_{j=1}^{r}\mathbf{u}_{jB}^T\boldsymbol{\Phi}_B\mathbf{u}_{jB}} \qquad (2)$$

The basis vectors of the subspace which minimize $\mathbf{F}_{com}$ are given by

$$(\boldsymbol{\Phi}_B^{-1}\boldsymbol{\Phi}_1)\mathbf{u}_j{=}\lambda_j\mathbf{u}_j \qquad (3)$$

where λ_j's are the smallest eigenvalues of the product $(\boldsymbol{\Phi}_B^{-1}\boldsymbol{\Phi}_1)$ and $\mathbf{u}_j$'s are the eigenvectors that correspond to these eigenvalues.

The metric $\mathbf{F}_{com}$ may solve the problem given in Fig.1. The common vectors of the two classes are shown in Fig.2 when the covariance matrix $(\boldsymbol{\Phi}_B^{-1}\boldsymbol{\Phi}_1)$ given in (3) is used. The derivation of the common vector for higher number of classes is similar with the two-class case, since the feature vectors of any class constitute one class and the feature vectors of the remaining classes can be considered as the feature vectors of the second class.

The common vector for all cases is determined by the subspace which minimizes $\mathbf{F}_{com}$ in (2), that is,

$$\mathbf{a}_{com_1} = \sum_{j=1}^{r}(\mathbf{a}_{ave_1}^T\mathbf{u}_j)\mathbf{u}_j$$

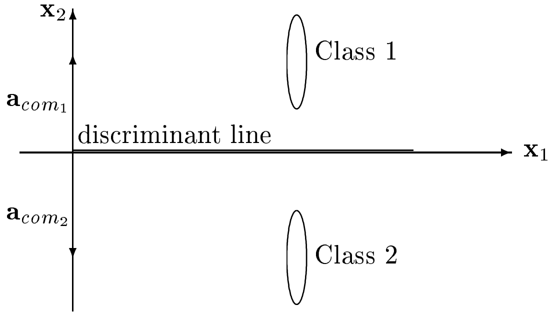

Figure 2. The common vectors are calculated in the difference subspace which is the direction of the $\mathbf{x}_2$ axis.

The discriminant line in Figure 2 may be drawn to pass through the average of the two commom vectors $\mathbf{a}_{com_1}$ and $\mathbf{a}_{com_2}$. Or it could be determined from the Bayes' theorem. One may note that this is a similar situation with the linear discriminant analysis (LDA) in a two-dimensional feature space. Since the Euclidean distance is usually used in the more than two-dimensional subspace of an d-dimensional feature space, the decision surface turns out to be a hypersphere within the subspace with its center pointed by the common vectors. A two-dimensional feature space with a one-dimensional subspace is shown in the following example.

Example 1: Let the points of the classes C_1 and C_2 be

$\mathbf{x}_1^1$=$[10 \quad 6]^T$ $\mathbf{x}_1^2$=$[10 \quad 10]^T$ $\mathbf{x}_1^3$=$[12 \quad 6]^T$ $\mathbf{x}_1^4$=$[12 \quad 10]^T$

$\mathbf{x}_2^1$=$[10 \quad -6]^T$ $\mathbf{x}_2^2$=$[10 \quad -10]^T$ $\mathbf{x}_2^3$=$[12 \quad -6]^T$

$\mathbf{x}_2^4$=$[12 \quad -10]^T$

The within-class and between-class covariance matrices of C_1 and C_2 are:

$$\mathbf{\Phi}_1 = \mathbf{\Phi}_2 = \begin{bmatrix} 4 & 0 \\ 0 & 16 \end{bmatrix} \qquad \mathbf{\Phi}_{B1} = \mathbf{\Phi}_{B2} = \begin{bmatrix} 4 & 0 \\ 0 & 1040 \end{bmatrix}$$

Then from (3),

$$\mathbf{\Phi}_B^{-1}{}_1\mathbf{\Phi}_1 = \mathbf{\Phi}_B^{-1}{}_2\mathbf{\Phi}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1/65 \end{bmatrix}$$

$\lambda_2 = \frac{1}{65}$ indicates the direction that minimizes $\mathbf{F}_{com}$. The basis vector for this subspace is then $\mathbf{u}_2$=$[0 \quad 1]^T$ which is the vertical direction. The common vector for the class C_1 and C_2 $\mathbf{a}_{com_1}$=$[0 \quad 8]^T$ and $\mathbf{a}_{com_2}$=$[0 \quad -8]^T$ respectively.

III. EXPERIMENTAL RESULTS WITH THE DECISION CRITERIA

For a feature vector $\mathbf{a}_x$ of an unknown class, one can easily obtain the remaining vector $\mathbf{a}_{x,rem_k}$:

$$\mathbf{a}_{x,rem_k} = \sum_{j=1}^{r} (\mathbf{a}_x^T \mathbf{u}_{j_k})\mathbf{u}_{j_k}$$

The Euclidean distance between the common vectors and the remaining vector is used as the decision criterion and it is given in the following:

$$\mathbf{C}^* = \underset{k}{\text{argmin}} ||\mathbf{a}_{x,rem_k} - \mathbf{a}_{com_k}|| \qquad (4)$$

If the feature vector $\mathbf{a}_x$ belongs to class C_k, then we expect to have a minimum distance between $\mathbf{a}_{x,rem_k}$ and $\mathbf{a}_{com_k}$.

In this study, 17 words which are numbers in Turkish "sıfır (zero), bir (one), iki (two), üç (three), dört (four), beç (five), altı (six), yedi (seven), sekiz (eight), dokuz (nine) " and Turkish words "aç (open), kapat (close), yanlış (false), evet (yes), hayır (no), ara (call), ofis (office)" are sampled at a rate of 9600 Hz with 12 bit accuracy [6]. This database yielded the same recognition rates with the TI-digit database in our previous work [1,2]. These words are taken from 200 speakers consisting of 127 male, 60 female, 13 children and they are stored in PC medium. Each spoken word can be represented as a vector. If the root-melcep parameters are derived for each frame with 256 samples of each word, the dimension of the feature vectors varied from 143 to 528 depending on the length of the spoken word. The experiment is continued by taking the first n=120 elements of the feature vectors so that every word in the database is represented by a (120x1) feature vector. The number of feature vectors p in each class in the training set is chosen to be p=150 for each class.

STUDY ON THE LDA

The criterion given in (1) is applied to the above database for the isolated word recognition and the results are given in Table 1. The recognition rate of 88% for 15 largest eigenvalues and of 85% for 12 largest eigenvalues are obtained for the training and test sets respectively. Since $\mathbf{S}_B$ in (1) has no inverse, Fisher's criterion can not be defined as

$$\mathbf{F}_{Fisher} = \frac{\mathbf{W}S_W\mathbf{W}^T}{\mathbf{W}S_B\mathbf{W}^T}. \qquad (5)$$

In order to use the criterion given above, the dimension of the feature vectors is reduced to 16 by using the eigenvectors of the between-class covariance matrix $\mathbf{S}_B$, since $\mathbf{S}_B$ has a dimension of 16x16 which is one less than the number of classes. The minimization of the criterion given in (5) yields the result that the weight values are determined by the eigenvectors corresponding to d' smallest eigenvalues of $\mathbf{S}_B^{-1}\mathbf{S}_W$. The definition of the Fisher's criterion as in (5) is more meaningful since $\mathbf{F}_{Fisher}$ has always a positive value. The minimum value of the metric $\mathbf{F}_{Fisher}$ can approach to zero as the best possible minimum value. However, the maximization of the metric may have a value which is too large and not meaningful for the recognition purposes.

Two cases are investigated by using the feature vectors with dimension reduced to 16.

a) The eigenvectors corresponding to the smallest eigenvalues of $\mathbf{S}_B^{-1}\mathbf{S}_W$ are used for the recognition purposes by using (4) and the results are given in Table 1.

b) In the second case, within-class covariance matrix $\mathbf{S}_W$ is calculated for each class $(\mathbf{S}_{W1}, \mathbf{S}_{W2}, ..., \mathbf{S}_{W17})$ instead of one covariance matrix as in the LDA. Therefore a different subspace for each class is determined from the corresponding within-class covariance matrix $\mathbf{S}_{Wk}$. Then two different scores S_k^1 and S^2 for the test vector are obtained for the $\mathbf{S}_{Wk}$ and $\mathbf{S}_B$ respectively. Final decision is given according to the minimum value of $||S_k^1/S^2||$ and the results are given in Table 1.

STUDY ON THE CVA

Five cases are investigated by using the CVA.

a) First of all, a subspace for each class C_k by using the metric $\mathbf{F}_{com}$ is determined from the eigenvectors of $\mathbf{\Phi}_{Bk}^{-1}\mathbf{\Phi}_{1k}$. When the eigenvectors corresponding to the smallest eigenvalues of $\mathbf{\Phi}_{Bk}^{-1}\mathbf{\Phi}_{1k}$ are used in the decision criterion (4), the recognition rates obtained for the training and test sets are given in Table 2.

b) In the second case, a different subspace for each class C_k is only determined from $\mathbf{\Phi}_{1k}$. The eigenvectors corresponding to the smallest eigenvalues of $\mathbf{\Phi}_{1k}$ are used in (4) and the recognition rates obtained for the training and test sets are given in Table 2.

c) A different subspace for each class C_k is only determined from $\mathbf{\Phi}_{Bk}$. The eigenvectors corresponding to the largest eigenvalues of $\mathbf{\Phi}_{Bk}$ are used in the following criterion:

$$\mathbf{C}^* = \underset{k}{\text{argmax}}||\mathbf{a}_{x,rem_k} - \mathbf{a}_{com_k}||$$

The recognition rates for the training and test sets are also given in Table 2.

d) In the fourth case, one score S_k^1 is obtained from the Euclidean distance between $\mathbf{a}_{x,rem_k}$ and $\mathbf{a}_{com_k}$ by using the eigenvectors of $\mathbf{\Phi}_{1k}$ and another score S_k^2 is obtained from the eigenvectors of $\mathbf{\Phi}_{Bk}$ for each class C_k. Then final decision is given according to the minimum value of $||S_k^1/S_k^2||$. The recognition rates are also given in Table 2.

e) In the last case, finally the scores S_k^1 and S_k^2 obtained in part (d) are combined to define the following criterion:

$$C^* = \text{index}\left\{ \min_{1 \le k \le 17} \{marked(S_k^1) + marked(S_k^2)\} \right\}. \quad (6)$$

Table 1. Recognition rates in the average obtained by using the LDA as percentage

	Training Set			Test Set		
λ^*	$S_W^{-1}S_B$	$S_B^{-1}S_W$	$\|\|S_k^1/S^2\|\|$	$S_W^{-1}S_B$	$S_B^{-1}S_W$	$\|\|S_k^1/S^2\|\|$
1	26	26	25	26	26	23
2	49	54	50	49	54	43
3	64	61	66	60	60	60
4	72	69	76	70	68	69
5	77	71	80	76	73	75
6	78	73	82	77	74	76
7	81	77	86	79	74	78
8	82	77	86	81	75	81
9	85	78	87	83	77	81
10	84	79	89	83	78	83
11	84	81	89	83	80	85
12	87	81	88	85	80	87
13	86	82	86	85	80	86
14	87	82	84	84	81	83
15	88	82	81	84	80	80
16	88	81	77	84	80	75

* The values of λ denotes the number of smallest eigenvalues for the minimization problems and the number of largest eigenvalues for the maximization problems.

Table 2. Recognition rates in the average obtained by using the CVA as percentage

	Training Set					Test Set				
λ^*	*a*	*b*	*c*	*d*	*e*	*a*	*b*	*c*	*d*	*e*
1	69	49	73	66	85	32	12	72	26	49
3	77	95	86	98	98	41	29	86	50	69
5	68	99	91	99	98	40	39	91	60	80
6	66	99	91	100	98	41	43	92	62	82
7	66	100	93	100	98	41	48	92	68	85
8	62	100	93	100	99	40	51	93	70	85
10	59	100	94	100	98	40	58	92	72	87
20	57	100	95	100	99	45	78	90	83	92
30	58	100	95	100	99	48	83	87	85	92
39	60	100	92	100	99	50	86	85	87	93
50	60	100	90	100	98	53	87	79	88	91
60	62	100	87	100	97	52	90	76	91	91
70	63	100	82	100	96	54	92	72	92	91
80	62	99	76	99	94	53	92	68	92	91
90	64	98	70	98	92	56	93	63	93	88
94	64	98	67	98	90	56	94	61	94	88
110	65	96	58	96	85	58	93	56	93	86
120	65	77	56	76	74	56	75	54	76	74

* The values of λ denotes the number of smallest eigenvalues for the minimization problems and the number of largest eigenvalues for the maximization problems.

In this criterion, the scores S_k^1 obtained for 17 classes are marked from the minimum to maximum and the scores and S_k^2 obtained for 17 classes are marked from the maximum to minimum. Then final decision is given according to (6) and the results are given in column (e) of the Table 2.

IV. CONCLUSION

When the common vector for multi-class case (metric $\mathbf{F}_{com}$) is compared with the Fisher's LDA, the classification rates of the CVA is superior to that of Fisher's LDA as seen from the comparison of the Tables 1 and 2. Especially in the training set, the CVA gives 100% recognition rate for 6 smallest eigenvalues. However, the LDA gives 89% as the maximum recognition rate in the training set.

As seen from the Table 1, the best recognition rates of 89% for the training set and of 87% for the test set are obtained in part (b) of the LDA with the feature vectors whose dimensions are reduced to 16 according to the between-class covariance matrix $\mathbf{S}_B$.

In the CVA, the maximum recognition rate of 94% is obtained in parts (b) and (d) for 94 smallest eigenvalues. In part (c) of the CVA, the maximum recognition rate of 93% is obtained for 8 largest eigenvalues. If the scores obtained for 17 classes in part (b) for 94 smallest eigenvalues and the scores obtained in part (c) for 8 largest eigenvalues are combined as in the parts (d) and (e) of the CVA, the recognition rates of 95% and 96% for the test set are obtained for parts (d) and (e) respectively and these results are not given in the Table 2.

The work on defining a more useful metric for the multi-class case is continuing.

REFERENCES

1. M. B. Gülmezoğlu, V. Dzhafarov, M. Keskin, and A. Barkana, “A novel approach to isolated word recognition,” IEEE Trans. on Speech and Audio Processing, vol. 7, no. 6, pp. 620-628, 1999.
2. M. B. Gülmezoğlu, V. Dzhafarov and A. Barkana, “The common vector approach and its relation to principal component analysis,” Accepted to be published in IEEE Trans. on Speech and Audio Processing, 2001.
3. M. Keskin, M. B. Gülmezoğlu, O. Parlaktuna and A. Barkana, “Isolated word recognition by extracting personal differences”, 6^{th} International Conference on Signal Processing Applications and Technology, Boston, USA, pp. 1989-1992, October 1995.
4. M. B. Gülmezoğlu and A. Barkana, “Text-Dependent Speaker Recognition by Using Gram-Schmidt Orthogonalization Method”, IASTED International Conference on Signal Processing and Communications, Canaria Islands, Spain, pp. 438-440, February 1998.
5. C. M. Bishop, Neural Networks for Pattern Recognition. Clarendon Press, Oxford, 1995.
6. A. Barkana, M. B. Gülmezoğlu, R. Edizkan, Ü. Künkçü, Ş. E. Künkçü, “Recognition of limited number of words by computer,” The Scientific and Technical Research Council of Turkey, Project Report, No. EEEAG-82, 1995.