Representative Class Vector Clustering-based Discriminant Analysis

Alexandros Iosifidis, Anastasios Tefas and Ioannis Pitas
Department of Informatics, Aristotle University of Thessaloniki, Greece
{*aiosif,tefas,pitas*}*@aiia.csd.auth.gr*

***Abstract*—Clustering-based Discriminant Analysis (CDA) is a well-known technique for supervised feature extraction and dimensionality reduction. CDA determines an optimal discriminant subspace for linear data projection based on the assumptions of normal subclass distributions and subclass representation by using the mean subclass vector. However, in several cases, there might be other subclass representative vectors that could be more discriminative, compared to the mean subclass vectors. In this paper we propose an optimization scheme aiming at determining the optimal subclass representation for CDA-based data projection. The proposed optimization scheme has been evaluated on standard classification problems, as well as on two publicly available human action recognition databases providing enhanced class discrimination, compared to the standard CDA approach.**

***Keywords*-Discriminant Analysis; feature selection; data projection; class representation**

I. Introduction

Clustering-based Discriminant Analysis (CDA) is a well known technique for supervised feature extraction and dimensionality reduction. Taking into account the subclass information, it determines a reduced dimensionality feature space, where samples belonging to different classes should be as far from another, while they should be as close as possible from the corresponding subclass center. It can be considered as a generalization of Linear Discriminant Analysis (LDA) since, by allowing multiple subclasses within each class, class multimodality is appropriately addressed. The adopted criterion is the ratio of the within-subclass scatter to the between-subclass scatter in the reduced dimensionality space. By minimizing this criterion, maximal class discrimination is achieved. CDA optimality is based on the assumptions that:

- all subclasses follow normal distributions having the same covariance structure and
- each subclass is represented by the corresponding subclass vector.

Although relying on rather strong assumptions that do not hold in many classification problems, it has been used in many applications, such as facial expression [1] and human action [2] recognition.

Taking into account that CDA optimization process encodes relationships between the class data by employing the corresponding subclass scatter matrices and by observing that such scatter matrices are functions of the corresponding subclass representative vectors, one may think that there could be several subclass representative vectors, other than the subclass mean, that could provide different scatter matrices enhancing class discrimination in the resulted reduced dimensionality space. In this paper, we propose an iterative optimization scheme aiming at determining such subclass representative vectors for CDA-based data projection.

The rest of the paper is structured as follows. In Section II, we briefly describe standard CDA algorithm. The proposed iterative optimization scheme aiming at determining the optimal subclass representative vectors is described in Section III. An experimental study on standard classification problems and publicly available action recognition datasests is provided in Section IV. Finally, conclusions are drawn in Section V.

II. Standard CDA

Given a set of D-dimensional data belonging to C classes $\mathbf{x}_{ijk} \in \mathbb{R}^D$, $i = 1, \ldots, C$, $j = 1, \ldots, c_i$, $k = 1, \ldots, N_{ij}$, where it is assumed that class i is formed by c_i subclasses, each containing N_{ij} samples., and their class labels $l_{ijk} = i$, standard CDA determines a projection matrix $\mathbf{W} \in \mathbb{R}^{D \times d}$, such that $\mathbf{y}_{ijk} = \mathbf{W}^T \mathbf{x}_{ijk}$ is the image of $\mathbf{x}_{ijk}$ in a d-dimensional feature space of increased class discriminative ability. The optimal projection matrix $\mathbf{W}^*$ is obtained by minimizing the ratio of the within-subclass scatter matrix $\mathbf{S}_w$ to the between-subclass scatter matrix $\mathbf{S}_b$ in the projection space. $\mathbf{S}_w$, $\mathbf{S}_b$ are determined by:

$$\mathbf{S}_w = \sum_{i=1}^{C} \sum_{j=1}^{c_i} \sum_{k=1}^{N_{ij}} \frac{1}{N_{ij}} (\mathbf{y}_{ijk} - \mathbf{m}_{ij})(\mathbf{y}_{ijk} - \mathbf{m}_{ij})^T, \quad (1)$$

$$\mathbf{S}_b = \sum_{i=1}^{C} \sum_{l \neq i} \sum_{j=1}^{c_i} \sum_{h=1}^{c_l} \frac{1}{C} (\mathbf{m}_{ij} - \mathbf{m}_{lh})(\mathbf{m}_{ij} - \mathbf{m}_{lh})^T, \quad (2)$$

where $\mathbf{m}_{ij} = \frac{1}{N_{ij}} \sum_{k=1}^{N_{ij}} \mathbf{y}_{ijk}$ is the mean vector of j-th subclass belonging to class i in the reduced dimensionality space $\mathbb{R}^d$. Since $\mathbf{y}_{ijk}$ are not a-priori known, it is convenient to express $\mathbf{S}_w$, $\mathbf{S}_b$ by using $\mathbf{x}_{ijk}$. It can be shown that:

$$\mathbf{S}_w = \mathbf{W}^T \bar{\mathbf{S}}_w \mathbf{W}, \quad (3)$$

and

$$\mathbf{S}_b = \mathbf{W}^T \bar{\mathbf{S}}_b \mathbf{W}. \quad (4)$$

$\bar{\mathbf{S}}_w$, $\bar{\mathbf{S}}_b$ are given by:

$$\bar{\mathbf{S}}_w = \sum_{i=1}^{C}\sum_{j=1}^{c_i}\sum_{k=1}^{N_{ij}} \frac{1}{N_{ij}}(\mathbf{x}_{ijk} - \boldsymbol{\mu}_{ij})(\mathbf{x}_{ijk} - \boldsymbol{\mu}_{ij})^T, \quad (5)$$

$$\bar{\mathbf{S}}_b = \sum_{i=1}^{C}\sum_{l\neq i}\sum_{j=1}^{c_i}\sum_{h=1}^{c_l} \frac{1}{C}(\boldsymbol{\mu}_{ij} - \boldsymbol{\mu}_{lh})(\boldsymbol{\mu}_{ij} - \boldsymbol{\mu}_{lh})^T. \quad (6)$$

where $\boldsymbol{\mu}_{ij} = \frac{1}{N_{ij}}\sum_{k=1}^{N_{ij}} \mathbf{x}_{ijk}$ is the mean vector of j-th subclass belonging to class i in the input space $\mathbb{R}^D$.

After obtaining $\bar{\mathbf{S}}_w$, $\bar{\mathbf{S}}_b$, $\mathbf{W}^*$ is calculated by solving the trace ratio optimization problem [3]:

$$\begin{aligned} \mathbf{W}^* &= \underset{\mathbf{W}^T\mathbf{W}=\mathbf{I}}{argmin}\ \mathcal{J}(\mathbf{W}), & (7) \\ \mathcal{J}(\mathbf{W}) &= \frac{Tr(\mathbf{W}^T\bar{\mathbf{S}}_w\mathbf{W})}{Tr(\mathbf{W}^T\bar{\mathbf{S}}_b\mathbf{W})}, & (8) \end{aligned}$$

where $Tr(\mathbf{A})$ denotes the trace of matrix $\mathbf{A}$, while the constraint $\mathbf{W}^T\mathbf{W} = \mathbf{I}$, $\mathbf{I} \in \mathbb{R}^{d\times d}$ being the identity matrix, is conventionally added to obtain a set of orthogonal and normalized projection vectors.

Since the trace ratio problem does not have a direct closed-form globally optimal solution [3], it is conventionally approximated by solving the ratio trace problem $\mathcal{J} = Tr[\left(\mathbf{W}^T\bar{\mathbf{S}}_b\mathbf{W}\right)^{-1}\left(\mathbf{W}^T\bar{\mathbf{S}}_w\mathbf{W}\right)]$, which is equivalent to the optimization problem $\bar{\mathbf{S}}_w\mathbf{v} = \lambda\bar{\mathbf{S}}_b\mathbf{v}$, $\lambda \neq 0$ and can be solved by applying eigenanalysis to the matrix $\bar{\mathbf{S}}_w^{-1}\bar{\mathbf{S}}_b$ in the case where $\bar{\mathbf{S}}_w$ is invertible, or to the matrix $\bar{\mathbf{S}}_b^{-1}\bar{\mathbf{S}}_w$ in the case where $\bar{\mathbf{S}}_b$ is invertible. $\mathbf{W}^*$ is formed by the eigenvectors corresponding to the non-zero eigenvalues. That is, the dimensionality of the resulted subspace is up to $d = \sum_{i=1}^{C} c_i$, since one may choose to keep fewer eigenvectors in order to form $\mathbf{W}^*$.

III. Proposed Optimization Scheme

As has been described in the previous Section, in CDA, class discrimination in the resulted low-dimensional space is measured by using the trace ratio value (8). By observing that $\bar{\mathbf{S}}_w$, $\bar{\mathbf{S}}_b$ are functions of $\boldsymbol{\mu}_{ij}$, as detailed in (5) and (6), respectively, we argue that there might be other subclass representative vectors that could increase class discrimination. Such subclass representative vectors can be obtained by minimizing CDA optimization criterion with respect to both the data projection matrix $\mathbf{W}$ and the subclass representative vectors $\tilde{\boldsymbol{\mu}}_{ij}$. That is, we propose to minimize the following criterion:

$$\tilde{\mathcal{J}}(\mathbf{W}, \tilde{\boldsymbol{\mu}}_{ij}) = \frac{Tr(\mathbf{W}^T\tilde{\mathbf{S}}_w(\tilde{\boldsymbol{\mu}}_{ij})\mathbf{W})}{Tr(\mathbf{W}^T\tilde{\mathbf{S}}_b(\tilde{\boldsymbol{\mu}}_{ij})\mathbf{W})}, \quad (9)$$

where the adopted subclass scatter matrices $\tilde{\mathbf{S}}_w$, $\tilde{\mathbf{S}}_b$ are determined by:

$$\tilde{\mathbf{S}}_w = \sum_{i=1}^{C}\sum_{j=1}^{c_i}\sum_{k=1}^{N_{ij}}(\mathbf{x}_{ijk} - \tilde{\boldsymbol{\mu}}_{ij})(\mathbf{x}_{ijk} - \tilde{\boldsymbol{\mu}}_{ij})^T, \quad (10)$$

$$\tilde{\mathbf{S}}_b = \sum_{i=1}^{C}\sum_{l\neq i}\sum_{j=1}^{c_i}\sum_{h=1}^{c_l}(\tilde{\boldsymbol{\mu}}_{ij} - \tilde{\boldsymbol{\mu}}_{lh})(\tilde{\boldsymbol{\mu}}_{ij} - \tilde{\boldsymbol{\mu}}_{lh})^T. \quad (11)$$

Since a simultaneous minimization of $\tilde{\mathcal{J}}$ with respect to both $\mathbf{W}$ and $\tilde{\boldsymbol{\mu}}_{ij}$ is not tractable, we propose an iterative optimization scheme consisting of two processing steps.

Let us denote by $\tilde{\boldsymbol{\mu}}_{ij,t}$ the subclass representative vectors determined for the t-th iteration of the proposed optimization scheme. Here we have introduced the index t denoting the iteration step of the proposed optimization scheme. $\tilde{\boldsymbol{\mu}}_{ij,t}$ are employed in order to determine the corresponding scatter matrices $\tilde{\mathbf{S}}_w(\tilde{\boldsymbol{\mu}}_{ij,t})$, $\tilde{\mathbf{S}}_b(\tilde{\boldsymbol{\mu}}_{ij,t})$ by using (10), (11), respectively. $\tilde{\mathbf{S}}_w(\tilde{\boldsymbol{\mu}}_{ij,t})$, $\tilde{\mathbf{S}}_b(\tilde{\boldsymbol{\mu}}_{ij,t})$ are, in turn, employed in order to determine the optimal data projection matrix $\mathbf{W}_t^*$ by solving the ratio trace problem.

After determining the optimal data projection matrix $\mathbf{W}_t^*$, $\tilde{\boldsymbol{\mu}}_{ij,t}$ are updated by following the direction of the gradient of (9), i.e.:

$$\tilde{\boldsymbol{\mu}}_{ij,t+1} = \tilde{\boldsymbol{\mu}}_{ij,t} - \beta\frac{\partial\tilde{\mathcal{J}}}{\partial\tilde{\boldsymbol{\mu}}_{ij,t}}. \quad (12)$$

The gradient $\frac{\partial\tilde{\mathcal{J}}}{\partial\tilde{\boldsymbol{\mu}}_{ij,t}}$ is given by:

$$\begin{aligned} \frac{\partial\tilde{\mathcal{J}}}{\partial\tilde{\boldsymbol{\mu}}_{ij,t}} &= \lambda_1\mathbf{W}_t^*\mathbf{W}_t^{*T}\left[\frac{2}{N_{ij}}\sum_{k=1}^{N_{ij}}\left(\tilde{\boldsymbol{\mu}}_{ij,t} - \mathbf{x}_{ijk}\right)\right] \\ &- \lambda_2\left[\mathbf{W}_t^*\mathbf{W}_t^{*T}\left(\frac{2}{C}\sum_{l\neq i}\sum_{h=1}^{c_l}\left(\tilde{\boldsymbol{\mu}}_{ij,t} - \tilde{\boldsymbol{\mu}}_{lh,t}\right)\right)\right], \end{aligned}$$

where:

$$\lambda_1 = \frac{1}{trace(\mathbf{W}_t^{*T}\tilde{\mathbf{S}}_b(\tilde{\boldsymbol{\mu}}_{ij,t})\mathbf{W}_t^*)}, \quad (13)$$

$$\lambda_2 = \frac{trace(\mathbf{W}_t^{*T}\tilde{\mathbf{S}}_w(\tilde{\boldsymbol{\mu}}_{ij,t})\mathbf{W}_t^*)}{trace(\mathbf{W}_t^{*T}\tilde{\mathbf{S}}_b(\tilde{\boldsymbol{\mu}}_{ij,t})\mathbf{W}_t^*)^2}. \quad (14)$$

β in (12) is an update rate parameter. In our experiments the value of β has been dynamically determined by using a line search strategy. That is, in each iteration of the proposed optimization scheme, the trace ratio criterion (9) has been evaluated by using an update rate parameter value equal to $\beta_0 = 0.1$. In the case where $\tilde{\mathcal{J}}(t+1) < \tilde{\mathcal{J}}(t)$ the trace ratio criterion has been evaluated by using an update rate parameter value $\beta_n = 2\beta_{n-1}$. This process has been followed until $\tilde{\mathcal{J}}(t+1) > \tilde{\mathcal{J}}(t)$ and the update rate parameter value providing the highest $\tilde{\mathcal{J}}$ decrease has been employed for subclass representative vectors adaptation. In the case where, by using an update rate parameter value equal to $\beta_0 = 0.1$, $\tilde{\mathcal{J}}(t+1) > \tilde{\mathcal{J}}(t)$, the trace ratio criterion has been evaluated by using an update rate parameter value $\beta_n = \beta_{n-1}/2$. This process has been followed until $\tilde{\mathcal{J}}(t+1) < \tilde{\mathcal{J}}(t)$ and the update rate parameter value providing $\tilde{\mathcal{J}}$ decrease has been employed for subclass representative vectors adaptation.

The above described procedure is applied until $\frac{\tilde{\mathcal{J}}(t)-\tilde{\mathcal{J}}(t+1)}{\tilde{\mathcal{J}}(t)} < \epsilon$, where ϵ is a small positive value. In our experiments we have used a value $\epsilon = 10^{-4}$.

IV. Experimental Results

In this Section, we present experiments conducted in order to evaluate the proposed optimization scheme. Since CDA is usually employed for feature selection and classification, we evaluated the performance of the proposed Representative Class Vector CDA (RCV-CDA) algorithm on standard classification problems. Furthermore, we have applied the proposed RCV-CDA algorithm on two publicly available human action recognition data sets. Details on the data sets used in our experiments are provided in the following subsections. In all the experiments, we compare the performance of the proposed RCV-CDA algorithm with that of standard CDA approach. In both cases, after data projection in the reduced dimensionality discriminant space, classification is performed by applying a modification of the nearest subclass centroid classification scheme. That is, a test sample $\mathbf{x}_t \in \mathbb{R}^D$ is mapped to the discriminant subspace by applying $\mathbf{y}_t = \mathbf{W}^{*T}\mathbf{x}_t$ and is assigned to the class label of the nearest subclass representative vector using the Euclidean distance, i.e.:

$$l_t = \underset{i}{argmin} \|\mathbf{y}_t - \tilde{\mathbf{m}}_{ij}\|_2,\ i = 1,\dots,C,\ j = 1,\dots,c_i, \tag{15}$$

where $\tilde{\mathbf{m}}_{ij} = \mathbf{W}^{*T}\tilde{\boldsymbol{\mu}}_{ij}$ is the image of the subclass representative vector $\tilde{\boldsymbol{\mu}}_{ij}$ in the reduced dimensionality space. For both the competing algorithms, we have conducted multiple experiments using different numbers of subclasses $c_i = 1, 2, 3, 4$ and we report the experiment provided the best performance in each data set.

A. Experiments on Standard Classification Problems

We conducted experiments on publicly available classification data sets coming from the machine learning repository of University of California Irvine (UCI) [4]. Information concerning the data sets used in our experiments can be found in Table I.

Table I
UCI data sets used in our experiments

Data set	# classes	# dimensions	# samples
Australian	2	14	690
Heart	2	13	270
Hill	2	100	1212
Indians	2	8	768
Ionosphere	2	34	351
Iris	3	4	150
Libras	15	90	360
Optdigits	10	64	5620
Skin	2	3	245057
Tic Tac Toe	2	9	958
Vertebral2c	2	6	310
Wine	3	13	178

Since there is not a standard training-testing split in these data sets, we have performed the 5-fold cross validation procedure for the two competing classification schemes by using the same partitioning. That is, in one experiment, each data set has been randomly split in five sets. The algorithms have been trained by using four sets and evaluated on the fifth set. This process has been performed five times (folds), one for each evaluation set, in order to complete an experiment. The mean classification rate over all folds has been used in order to measure the performance of each algorithm in one experiment. Ten experiments have been performed in total and the mean classification rate over all experiments, as well as the standard deviation of the observed results, have been calculated in order to measure the performance of each algorithm in each data set. The results obtained for these experiments are illustrated in Table II.

Table II
Classification results on UCI data sets

Data set	CDA	RCV-CDA
Australian	74,86% (3,49%)	**75,7% (3,46%)**
Heart	53,67% (2,34%)	**63,04% (3,83%)**
Hill	57,18% (1,84%)	**58,34% (2,47%)**
Indians	56,26% (2,17%)	**57,55% (1,41%)**
Ionosphere	68,58% (4,98%)	**72,65% (2,83%)**
Iris	96,8% (0,69%)	**96,93% (0,64%)**
Libras	75,19% (1,39%)	**75,22% (1,53%)**
Optdigits	95,7% (0,14%)	**95,74% (0,25%)**
Skin	39,34% (3,32%)	**41,52% (2,85%)**
Tic Tac Toe	51,05% (1,84%)	**52,41% (1,22%)**
Vertebral2c	66,71% (3,3%)	**69,1% (3,68%)**
Wine	95,28% (0,66%)	**95,67% (0,53%)**

As can be seen in Table I, the classification scheme employing the proposed RCV-CDA algorithm outperforms the one employing the standard CDA algorithm in all cases, providing $1-9\%$ improvement on the performance of the standard CDA algorithm.

B. Experiments on Human Action Recognition

We have conducted experiments on two publicly available action recognition data sets, namely i3DPost and Hollywood2. Information concerning each data set is provided in the following.

1) The i3DPost *data set [5]:* consists of 64 sequences depicting eight persons performing eight actions. Eight cameras having a wide 45^o viewing angle difference to provide 360^o coverage of the scene were placed on a ring of 8m diameter at a height of 2m above the studio floor. The studio was covered by blue background. The actions appearing in the database are: 'walk', 'run', 'jump in place', 'jump forward', 'bend', 'fall', 'sit on a chair', and 'wave one hand'. Example video frames depicting a person in the database walking are illustrated in Figure 1. In our experiments, we have adopted the dyneme-based video representation [7] and applied the Leave-One-Person-Out cross-validation scheme.

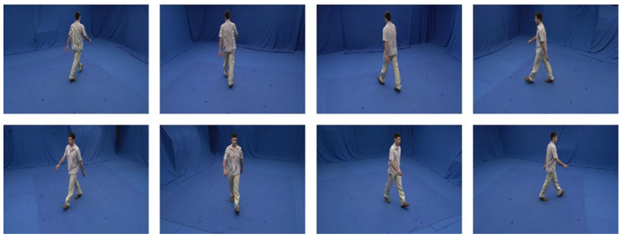

Figure 1. *Example video frames depicting a person walking from all the eight cameras.*

That is, the algorithms have been trained by using the videos depicting seven person in the database and tested on the videos depicting the remaining one. This process has been repeated eight times (folds), one for each test person. The mean action classification rate over all folds has been used in order to measure the performance of each competing algorithm.

2) The Hollywood2 data set [6]: consists of 1707 videos taken from 69 different Hollywood movies. The actions appearing in the dataset are: 'answering phone', 'driving car', 'eating', 'fighting', 'getting out of the car', 'hand shaking', 'hanging', 'kissing', 'running', 'sitting down', 'sitting up' and 'standing'. Example video frames are illustrated in Figure 2. In our experiments we have adopted the data partitioning provided in the database, i.e., 823 videos were used for training the algorithms and 884 videos have been used for evaluation. We have adopted the Bag of Visual Words (BoVWs)-based video representation employing the Histogram of Oriented Gradients (HOG) and the Histograms of Optical Flow (HOF) descriptors calculated on Space-Time Interest Points [8].

The action classification rates obtained for both the proposed RCV-CDA and the standard CDA-based classification schemes are illustrated in Table III. As can be seen in this Table, the proposed RCV-CDA-based classification scheme outperforms the one based on standard CDA algorithm in both cases, providing 3% and 2% improvement on the i3DPost and the Hollywood2 data sets, respectively.

Table III
CLASSIFICATION RESULTS ON THE I3DPOST AND HOLLYWOOD2 DATA SETS

Data set	CDA	RCV-CDA
i3DPost	90.63%	**93.75%**
Hollywood2	38.35%	**40,50%**

V. CONCLUSIONS

In this paper, we proposed an optimization scheme aiming at the optimal class representation for CDA-based data projection. This has been done by optimizing the CDA criterion with respect to both the data projection matrix and the subclass representative vectors following an iterative optimization scheme. The proposed RCV-CDA algorithm has been evaluated on standard classification problems, as well as on two publicly available action recognition data sets providing enhanced classification performance, compared to the standard CDA approach.

Figure 2. *Example video frames depicting all the 12 action classes of the Hollywood2 database.*

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 316564 (IMPART). This publication reflects only the authors views. The European Union is not liable for any use that may be made of the information contained therein.

REFERENCES

[1] X. Chen and T. Huang, Facial expression recognition: a clustering-based approach, Pattern Recognition Letters, vol. 24, no. 10, pp. 1295-1302, 2003.

[2] A. Iosifidis, A. Tefas and I. Pitas, Multi-view action recognition based on action volumes, fuzzy distances and cluster discriminant analysis, Signal Processing, vol. 93, no. 6, pp. 1445-1457, 2013.

[3] Y. Jia, F. Nie and C. Zhang, Trace ratio problem revisited, IEEE Transactions on Neural Networks, vol. 20, no. 4, pp. 729-735, 2009.

[4] K. Bache and M. Lichman, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2013.

[5] N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis and I. Pitas, The i3DPost multi-view and 3D human action/interaction database, Conference on Visual Media Production, 2009.

[6] M. Marszalek, I. Laptev and C. Scmid, Actions in context, Computer Vision and Pattern Recognition, 2009.

[7] A. Iosifidis, A. Tefas and I. Pitas, View-Invariant Action Recognition Based on Artificial Neural Networks, IEEE Transactions on Neural Networks and Learning Systems , vol. 23, no. 3, pp. 412–424, 2012.

[8] H. Wang, M. M. Ullah, A. Klser, I. Laptev and C. Schmid, Evaluation of Local Spatio-temporal Features for Action Recognition. British Machine Vision Conference, 2009.