Canonical Correlation Analysis for Multiview Semisupervised Feature Extraction

Olcay Kursun[1] and Ethem Alpaydin[2]

[1] Department of Computer Engineering, Istanbul University, 34320, Avcilar, Istanbul, Turkey
okursun@istanbul.edu.tr

[2] Department of Computer Engineering, Bogazici University, 34342, Bebek, Istanbul, Turkey
alpaydin@boun.edu.tr

Abstract. Hotelling's Canonical Correlation Analysis (CCA) works with two sets of related variables, also called views, and its goal is to find their linear projections with maximal mutual correlation. CCA is most suitable for unsupervised feature extraction when given two views but it has been also long known that in supervised learning when there is only a single view of data given, the supervision signal (class-labels) can be given to CCA as the second view and CCA simply reduces to Fisher's Linear Discriminant Analysis (LDA). However, it is unclear how to use this equivalence for extracting features from multiview data in semisupervised setting (i.e. what modification to the CCA mechanism could incorporate the class-labels along with the two views of the data when labels of some samples are unknown). In this paper, a CCA-based method supplemented by the essence of LDA is proposed for semi-supervised feature extraction from multiview data.

Keywords: Semisupervised Learning; Feature Extraction; Multiview Learning; LDA; CCA.

1 Introduction

Fisher's Linear Discriminant Analysis (LDA) [1] is one of the most popular linear dimensionality reduction methods; it seeks to find discriminatory projections of the data (i.e. those, which maximize the between class scatter and minimize the within class scatter). Whereas, Hotelling's Canonical Correlation Analysis (CCA) [2] works with two sets of related variables and its goal is to find maximally correlated linear projections of the two sets of variables. While LDA works completely in supervised setting (e.g. computationally, it needs to compute the within and between-class scatter matrices), CCA works completely in unsupervised manner (i.e. it ignores the class-labels and looks for correlated functions between the two views of data samples). Finding such correlated functions of the two views of the same phenomenon by discarding the representation-specific details (noise) is expected to reveal the underlying hidden yet influential semantic factors responsible for the correlation [3]. In this work, we extend the CCA setup so that it can take into account the class-label information into account as well. There are various ways of extending CCA to work with more than two views [4]; however, considering the

L. Rutkowski et al. (Eds.): ICAISC 2010, Part I, LNAI 6113, pp. 430-436, 2010.

class-label information as a third view is not directly applicable in the semisupervised setting.

We propose to accommodate the class-labels in the CCA setup in a rather indirect way, through the class centers of the other view. Thus, if all the samples were labelled, this setup reduces to the classical samples versus class-labels setup, which has been long known to be equivalent to LDA with the slight change of representation for the class-labels by representing them using the mean of the samples of that class in the other view rather than a kind of 1-of-C coding [5]. On the other hand, if all the samples were unlabelled this setup is the plain CCA itself. However, when there are both labelled and unlabelled samples, our method extracts CCA-like features with preference for LDA-like discriminatory ones.

2 Canonical Correlation Analysis (CCA)

Canonical Correlation Analysis (CCA) is introduced by Hotelling (1936) to describe the linear relations between two multidimensional (or two sets of) variables as the problem of finding basis vectors for each set such that the projections of the two variables on their respective basis vectors are maximally correlated (Figure 1). These two sets of variables, for example, may correspond to different views of the same semantic object (e.g. audio versus video of a person speaking, two cameras viewing the same object as in binocular vision, text versus links or images in webpages, etc). Let u-dimensional X and v-dimensional Y denote corresponding two sets of real-valued random variables (i.e., $X \in \mathbb{R}^u$ and $Y \in \mathbb{R}^v$), the canonical correlation is defined as:

$$\rho(X;Y) = \sup_{f,g} \ \mathrm{corr}(f^T X; g^T Y) \tag{1}$$

where, corr(X;Y) stands for Pearson's correlation coefficient.

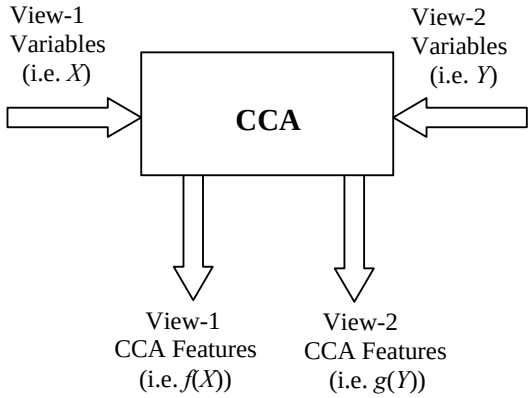

Fig. 1. CCA-based Feature Extraction. Correlated features are extracted from the two views. The class-labels are not utilized.

The problem of finding the orthogonal projections that achieve the top correlations reduces to a generalized eigenproblem, where the projection f (and the projection g can be solved for similarly) corresponds to the top eigenvector of the following [6]:

$$\mathbf{C}_{XX}^{-1}\mathbf{C}_{XY}\mathbf{C}_{YY}^{-1}\mathbf{C}_{YX}f = \lambda_{CCA}f \tag{2}$$

and

$$\rho(X;Y)=\sqrt{\lambda_{CCA}}\,, \tag{3}$$

where

$$\mathbf{C}(X,Y) = \mathrm{E}\left\{\begin{pmatrix} X \\ Y \end{pmatrix}\begin{pmatrix} X \\ Y \end{pmatrix}^T\right\} = \begin{bmatrix} \mathbf{C}_{XX} & \mathbf{C}_{XY} \\ \mathbf{C}_{YX} & \mathbf{C}_{YY} \end{bmatrix}. \tag{4}$$

3 Fisher Linear Discriminant Analysis (LDA)

Fisher Linear Discriminant Analysis (LDA) is a variance preserving approach with the goal of finding the optimal linear discriminant function [1, 7]. To utilize the categorical class label information in finding informative projections, LDA considers maximizing an objective function that involves the scatter properties of every class as well as the total scatter [7]. The objective function is designed to be maximized by a projection that maximizes the between class (or equivalently total scatter as in PCA) and minimize the within class scatter:

$$J = \sup\frac{h^T\mathbf{S}_B h}{h^T\mathbf{S}_W h}\,. \tag{5}$$

The optimization can be shown to be accomplished by computing the solution of the following generalized eigenproblem for the eigenvectors corresponding to the largest eigenvalues:

$$\mathbf{S}_B h = \lambda_{LDA}\mathbf{S}_W h\,. \tag{6}$$

LDA is originally designed for single view datasets, therefore when we have two views of the same objects (as the case for CCA), one straightforward approach would be to use the views separately as shown in Figure 2 and use both feature sets together for the subsequent classification task.

A direct connection between LDA and CCA can be obtained by showing that LDA is exactly what is accomplished by applying CCA between the set of all variables (of a view) and the corresponding class labels (0/1 for binary, 1-of-C coding for multiclass classification). Searching for the maximal correlations between the variables and the class-labels via CCA (Figure 3), yields the LDA projections as solutions [5, 8, 9].

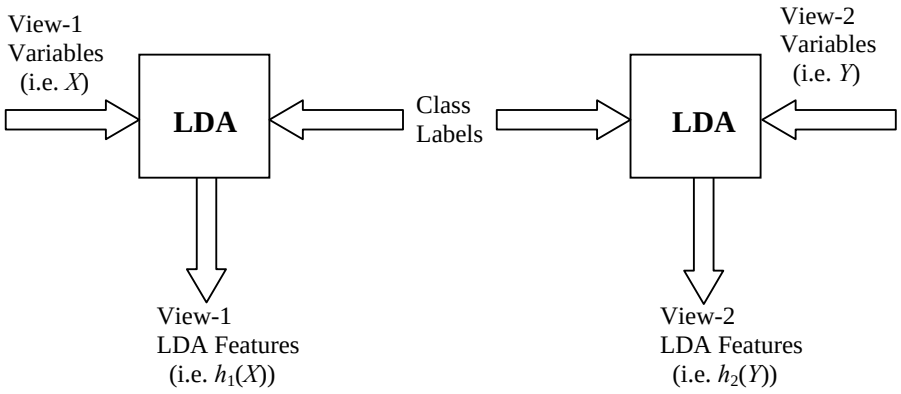

Fig. 2. LDA-based Feature Extraction. Features are extracted from the two views independently only for the labelled samples.

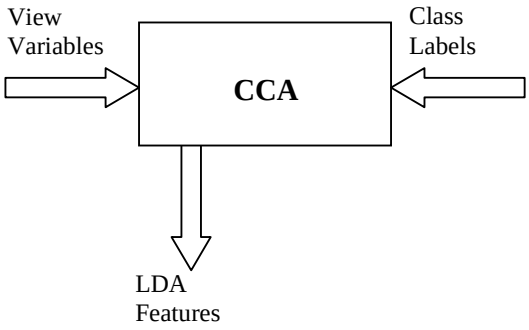

Fig. 3. CCA-based implementation of LDA. Correlated functions of the (single) view and the class-labels correspond to features extracted by LDA.

4 Proposed Architecture for Semisupervised CCA (SCCA)

One key observation to the method we propose is that in the architecture presented in Figure 3, the class-labels are not required to be hard labels in discrete format (e.g. class-0 and class-1 represented as 0 and 1 respectively). In fact, class-centers can be presented as class-labels [10]. Our proposal is simply to keep the other view when the class-label is absent; and otherwise, represent the class-labels by replacing the other view by the class-center of the samples in that other view. For example, to extract such SCCA features for View-1, we use View-1 variables versus View-2 variables in a regular CCA setup but we change View-2 feature vector to the respective class-center vector for the labelled samples (Figure 4). The procedure can be repeated in a similar fashion in order to extract SCCA features for View-2. Thus,

SCCA features are expected to represent the view to view relations (akin to CCA) as well as view to class relations (akin to LDA) because for the unlabelled samples SCCA works like CCA and for the labelled samples it works like LDA.

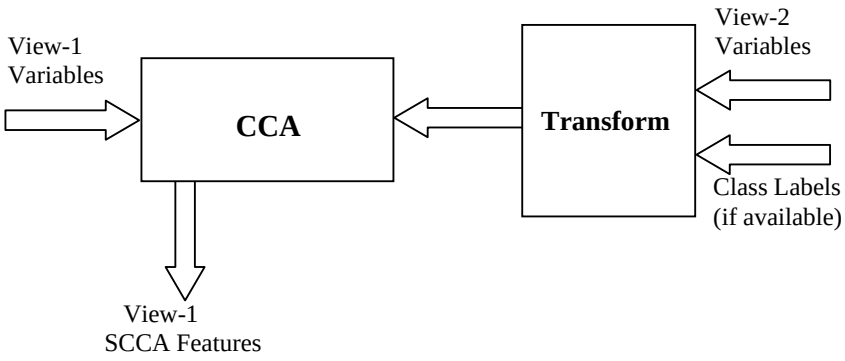

Fig. 4. The proposed semisupervised version of CCA-based Feature Extraction (View-2 features can be extracted similarly). When dealing with a labelled data sample, View-2 variables are replaced by the View-2 prototype (class center) of the class of that sample.

5 Experimental Results

For our experiments, we have used "Multi-feature digit dataset" [11] available from UCI machine learning repository [12]. This dataset consists of features of handwritten numerals (digits from '0' to '9') extracted from a collection of Dutch utility maps. 200 samples per class have been digitized and then represented in terms of the following six feature sets:

1. mfeat-fou: 76 Fourier coefficients of the character shapes;
2. mfeat-fac: 216 profile correlations;
3. mfeat-kar: 64 Karhunen-Loève coefficients;
4. mfeat-pix: 240 pixel averages in 2 x 3 windows;
5. mfeat-zer: 47 Zernike moments;
6. mfeat-mor: 6 morphological features.

Among the 200 samples per-class, we used the first 100 for the training (to account for both labelled and unlabelled) and the remaining 100 samples for the testing. We varied the number of labelled/unlabelled samples in the training set to evaluate the contribution of the unlabelled samples to plain LDA that only uses the labelled samples and also to evaluate the contribution of the labelled samples to plain CCA that uses all the available training samples but without benefiting from the class information of the labeled ones. We used CCA implementation in [13].

As some pairs of views can better complement weaknesses of each other than some others, we have avoided picking a particular pair of views; instead, we applied SCCA to all the 15 pairwise combinations of these six views. In Figure 5, we show the test accuracies averaged over 750 = 15×50 classification runs (15 pairs of views

and 50 random splits of the training set into labelled and unlabelled groups for each view-pair). We can see that for low ratio of labelled samples SCCA achieves the highest accuracy levels and LDA performs poorly. However, as the number of labelled samples increase relative to the unlabelled ones, LDA performs better because the use of unlabelled samples introduce noise and simply shifts the optimal decision boundary unnecesarily. For the training and testing we used LIBSVM [14] implementation of linear SVM-classifiers and as inputs to the SVM we extracted the same number of features from both views (shown as the title at the top of plots in each panel). The fact that we used linear SVM for classification shows that SCCA features are clearly superior to LDA and CCA features when there are abundance of unlabelled samples.

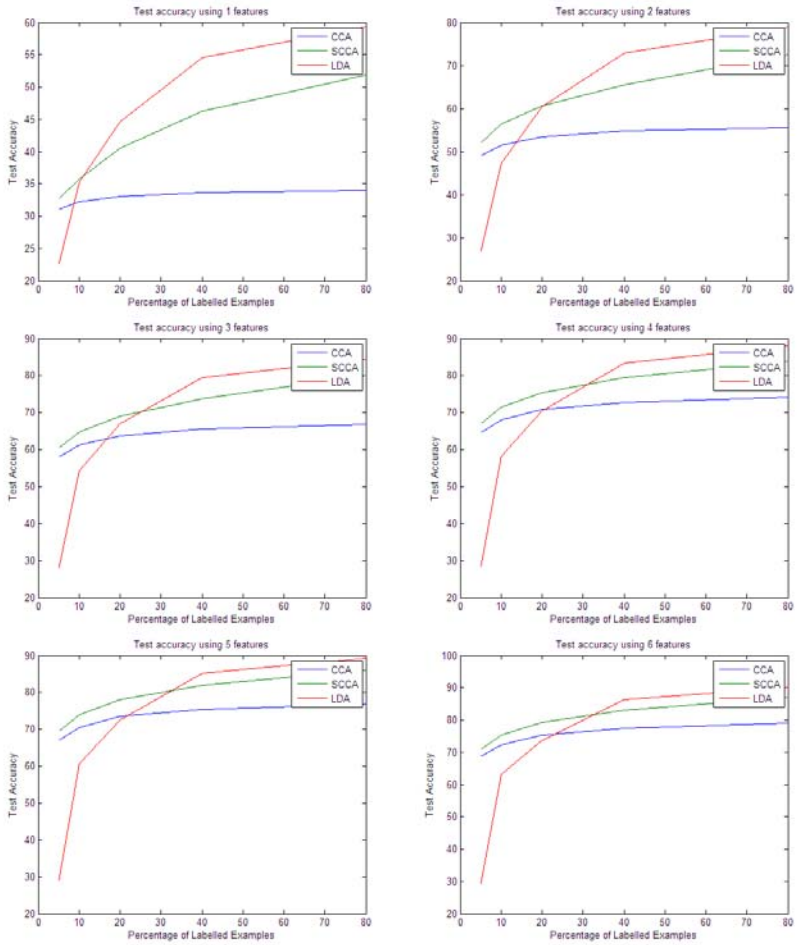

Fig. 5. SVM classification accuracies using various number of features extracted (per view) by CCA, LDA, and the proposed SCCA methods.

6 Conclusions

In this paper, we proposed a method called SCCA for semisupervised multiview feature extraction. We propose to use CCA with a modification to accommodate the class-labels through the class centers of the other view. Even though, we limited ourselves to two-view (plus the class-labels) scenario, the results can be generalized to more views [4]. To extract SCCA features of a view, we use that view and also the unlabelled samples of the other view as is; but we transform the labelled samples of the other view by replacing them with their corresponding class-centers in that (other) view. Thus, labelled samples are replaced by their prototypes and provide a form of LDA-like supervision to the proposed CCA-like setup. Thus, if all the samples were labelled, this setup reduces to LDA; and if all the samples were unlabelled, it simply is the plain CCA. However, when there are both labelled and unlabelled samples, our method extracts CCA-like features with preference for LDA-like discriminatory ones. The experimental results on a benchmark, multi-feature digit dataset, shows that SCCA features are clearly more advantageous than both LDA and CCA features when the number of labelled samples are small and there are a large number of unlabelled ones.

References

1. Fisher, R.A. (1936) The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7: 179-188.
2. Hotelling, H. (1936) Relations between two sets of variates. *Biometrika* 28: 321-377.
3. Favorov, O.V., Ryder, D. (2004) SINBAD: a neocortical mechanism for discovering environmental variables and regularities hidden in sensory input. *Biological Cybernetics* 90: 191-202.
4. Kettenring, J.R. (1971) Canonical analysis of several sets of variables. *Biometrika* 58: 433-451.
5. Bartlett, M.S. (1938) Further aspects of the theory of multiple regression. *Proc. Camb. Philos. Soc.* 34: 33-40.
6. Hardoon, D., Szedmak, S., Shawe-Taylor, J. (2004) Canonical correlation analysis: an overview with application to learning methods. *Neural Computation* 16: 2639-2664.
7. Alpaydin, E. (2004) Introduction to Machine Learning (Adaptive Computation and Machine Learning Series). The MIT Press.
8. Loog, M., van Ginneken, B., Duin, R.P.W. (2005) Dimensionality reduction of image features using the canonical contextual correlation projection. *Pattern Recognition* 38: 2409-2418.
9. Barker, M., Rayens, W. (2003) Partial least squares for discrimination. *Journal of Chemometrics* 17: 166-173.
10. Sun, T., Chen, S. (2007) Class label versus sample label-based CCA. *Applied Mathematics and Computation* 185: 272-283.
11. van Breukelen, M., Duin, R.P.W., Tax, D.M.J., den Hartog, J.E. (1998) Handwritten digit recognition by combined classifiers. *Kybernetika* 34(4): 381-386.
12. Asuncion, A., Newman, D.J., UCI Machine Learning Repository. Irvine, CA: University of California, Department of Information and Computer Science, 2007.
13. Borga, M. (1998) Learning Multidimensional signal processing, PhD thesis, Department of Electrical Engineering, Linköping University, Linköping, Sweden.
14. Hsu, C.W., Lin, C.J., A Comparison of Methods for Multi-Class Support Vector Machines. IEEE Trans. Neural Networks, vol. 13, 2002, pp. 415-425.