Supervised Cross-Modal Factor Analysis for Multiple Modal Data Classification

Supervised cross-modal factor analysis for multiple modal data classification

Jingbin Wang
National Time Service Center
Chinese Academy of Sciences, Xi' an 710600 , China
Graduate University of Chinese Academy of Sciences
Beijing 100039, China
jingbinwang1@outlook.com

Yihua Zhou
Department of Mechanical Engineering and Mechanics
Lehigh University
Bethlehem, PA 18015, USA

Kanghong Duan
North China Sea Marine Technical Support
Center, State Oceanic Administration
Qingdao 266033, China

Jim Jing-Yan Wang
Computer, Electrical and Mathematical
Sciences and Engineering Division
King Abdullah University
of Science and Technology
Thuwal 23955, Saudi Arabia

Halima Bensmail*
Qatar Computing Research Institute
Doha 5825, Qatar
*Corresponding author

***Abstract*—In this paper we study the problem of learning from multiple modal data for purpose of document classification. In this problem, each document is composed two different modals of data, i.e., an image and a text. Cross-modal factor analysis (CFA) has been proposed to project the two different modals of data to a shared data space, so that the classification of a image or a text can be performed directly in this space. A disadvantage of CFA is that it has ignored the supervision information. In this paper, we improve CFA by incorporating the supervision information to represent and classify both image and text modals of documents. We project both image and text data to a shared data space by factor analysis, and then train a class label predictor in the shared space to use the class label information. The factor analysis parameter and the predictor parameter are learned jointly by solving one single objective function. With this objective function, we minimize the distance between the projections of image and text of the same document, and the classification error of the projection measured by hinge loss function. The objective function is optimized by an alternate optimization strategy in an iterative algorithm. Experiments in two different multiple modal document data sets show the advantage of the proposed algorithm over other CFA methods.**

***Index Terms*—Multiple modal learning, Cross-modal factor analysis, Supervised learning**

I. Introduction

In this paper, we deal with the problem of learning from multiple modal data [1], [2]. Traditional data representation, classification and retrieval problems usually focus on single modal data [3], [4]. For example, for the problem of text classification, we usually only consider using a data set of text to train a classifier [5], [6]. While for the problem of image representation, only the images are considered to learn the representation parameters [7], [8]. However, in modern information landscape, the data is usually composed of d-ifferent modals. For example, in video clips, there are two modals of data, e.g. image sequence data and audio data. Moreover, in a research article, there are not only the text data, but also the image data. Learning from these multiple modal data has attracted much attention from both machine learning and multimedia information processing communities [9], [10]. Recently, cross-modal factor analysis (CFA) has been proposed to project different modal of data to a shared feature space so that classification or retrieval can be performed cross data modal [11]. It assumes that each document is composed of two modals of data, e.g., image and text, and try to learn a projection matrix for each modal, so that the projections of the two modals of a document can be as close to each other as possible. With these projections, we can project any data of one of the two modals to the shared data space, and then perform retrieval cross these two modals. Moreover, since the data of different modals can be projected to a shared data space, we can also train a classifier from both the modals for the classification of data of any modal. CFA is further extended to its kernel version in [12].

In this paper, we consider the problem of learning a classifier from multiple modal data, and using the classifier to classify a single modal data. In a training set of documents, each document is composed of data of two modals, e.g., an image and a text. We can first apply CFA to project two modals to a shared space and then learn a classifier in this space by using the class label information. However, in the phase of projection, the class label information has been ignored by CFA. Actually, without using supervision information provided by class label information, it cannot be guaranteed that the projections are discriminative enough. Although we can apply powerful classification methods after the projection of multiple modal data, the discriminative information which is necessary for classification may has been lost during the projection procedure. Thus it is very necessary to include the available

class label information in the CFA projection. Surprisingly, no work has been done to incorporate the supervision information contained by the class labels to improve the discriminative ability of CFA. To fill this gap, in this paper, we proposed the first supervised CFA method by regularizing the projections of different modals by the class label information.

The contributions of this paper are of two folds:

1) For the first time, we propose the formulation of supervised CFA. We propose to project the data of two modals to a shared data space by orthonormal transformations, and use a linear class label function to predict the class labels of the training multiple modal documents. To formulate the problem, we propose to minimize the difference between the projections of two modals of the same document, and simultaneously minimize the classification error of both modals measured by hinge loss. In this way, the learning of orthonormal transformation matrices of multiple modal projections and the class label predictor parameter are unified, and the predictor learning can regularize the learning of orthonormal transformation matrices to improve the discriminative ability of the multiple modal projections.
2) We also develop an iterative algorithm to optimize the constrained minimization problem of this formulation. The projection parameter and the predictor parameter are optimized alternately in an iterative algorithm. The orthonormal transformation matrices are optimized by fixing class label predictor parameter matrix and solving a singular value decomposition (SVD) problem [13], [14]. The class label predictor parameter matrix is solved by fixing orthonormal transformation matrices and solving a quadratic programming (QP) problem. [15], [16]

The remaining of this paper is organized as follows: in section II we introduce the proposed supervised CFA (SupCFA), in section III the proposed method is evaluated experimentally, and in section IV, the paper is concluded.

II. Proposed method

A. Problem formulation

We assume that we have a training set of n documents $\mathcal{D} = \{D_1, \cdots, D_n\}$, where D_i is the i-th document. Each document is comprised of an image component and an accompanying text, i.e., $D_i = (I_i, T_i)$, where $I_i \in \mathbb{R}^{d_I}$ is a d_I-dimensional feature space of the image, $T_i \in \mathbb{R}^{d_T}$ is a a d_T-dimensional feature of the text. These documents are assumed to belongs to m classes, and for the i-th document, we define a class label vector $\mathbf{y}_i = [y_{i1}, \cdots, y_{im}] \in \{+1, -1\}^m$ to indicate which class it belongs to. The j-th dimension of $\mathbf{y}_i$, $y_{ij} = +1$ if D_i belongs to the j-th class, and $y_{ij} = -1$ otherwise. The goal of cross-modal classification is to learn a predictor from the training set, and use the predictor to predict the class label vector of given text (image) query T (I).

To this end, we first project images and texts of documents to a shared d-dimensional feature space by orthonormal transformations, and then learn a linear prediction function in the share space to predict the class label vector. The projection in the image and text space are given as follows:

$$I_i \to I_i\Omega_I \in \mathbb{R}^d,\ T_i \to T_i\Omega_T \in \mathbb{R}^d \tag{1}$$

where $\Omega_I \in \mathbb{R}^{d_I \times d}$ is the orthonormal transformation matrix of the image data, and $\Omega_T \in \mathbb{R}^{d_T \times d}$ is the orthonormal transformation matrix of the text data. CFA assumes that the projections of the image and text of a single document should be as close to each other as possible, and the squared ℓ_2 norm distance between $I_i\Omega_I$ and $T_i\Omega_T$ is minimized over all the training documents,

$$\begin{aligned} \min_{\Omega_I, \Omega_T}\ & \sum_{i=1}^{n} \|\Omega_I I_i - \Omega_T T_i\|_2^2 \\ s.t.\ & \Omega_I^\top \Omega_I = I_{d\times d}, \Omega_T^\top \Omega_T = I_{d\times d}, \end{aligned} \tag{2}$$

where $I_{d\times d}$ is a $d_I \times d_I$ identity matrix. To predict the class label vector $\mathbf{y}_i$ from the projections of image and text of the i-th document, we try to learn a linear function as follows,

$$\begin{aligned} \mathbf{y}_i &\leftarrow f_W(I_i\Omega_I) = (I_i\Omega_I)W, \\ \mathbf{y}_i &\leftarrow f_W(T_i\Omega_T) = (\Omega_T T_i)W, \end{aligned} \tag{3}$$

where $W \in \mathbb{R}^{d\times m}$ is the predictor parameter matrix. To learn W, we minimize its squared ℓ_2 norm and the hinge loss of the predictor over both the images and texts of all the training documents,

$$\begin{aligned} \min_{W}\ & \|W\|_2^2 + C_1 \sum_{i=1}^{n} (\xi_i + \varepsilon_i) \\ s.t.\ & h - (I_i\Omega_I)W\mathbf{y}_i^\top \le \xi_i, \xi_i \ge 0, \\ & h - (\Omega_T T_i)W\mathbf{y}_i^\top \le \varepsilon_i, \varepsilon_i \ge 0, i = 1, \cdots, n. \end{aligned} \tag{4}$$

where ξ_i is the slack variable of the hinge loss over the image of the i-th document, ε_i is that of the text of the i-th document, h is a parameter of the hinge loss function, and C_1 is a tradeoff parameter. The minimization of $\|W\|_2^2$ is to reduce the complexity of the predictor and also to seek a large margin. The hinge loss is applied to reduce the prediction error.

The formulation of the proposed method is the combination of (2) and (4), which is as follows,

$$\begin{aligned} \min_{W,\Omega_I,\Omega_T}\ & \frac{1}{2}\|W\|_2^2 + C_1 \sum_{i=1}^{n} (\xi_i + \varepsilon_i) + C_2 \sum_{i=1}^{n} \|I_i\Omega_I - T_i\Omega_T\|_2^2 \\ s.t.\ & h - (I_i\Omega_I)W\mathbf{y}_i^\top \le \xi_i, \xi_i \ge 0, \\ & h - (T_i\Omega_T)W\mathbf{y}_i^\top \le \varepsilon_i, \varepsilon_i \ge 0, i = 1, \cdots, n, \\ & \Omega_I^\top \Omega_I = I_{d\times d}, \Omega_T^\top \Omega_T = I_{d\times d}, \end{aligned} \tag{5}$$

where C_2 is another tradeoff parameter. Please note that in this formulation, not only the orthonormal transformation matrices are variables, but also the predictor parameter. In this way, the representation and the classification of image and text modals are unified. Both the images and texts are mapped to a shared space and then a shared predictor are applied to classify them.

B. Optimization

To solve the problem in (5), we write the Lagrange function as follows,

$$\begin{aligned}\mathcal{L} =&\frac{1}{2}\|W\|_2^2 + C_1\sum_{i=1}^{n}(\xi_i+\varepsilon_i) + C_2\sum_{i=1}^{n}\|I_i\Omega_I - T_i\Omega_T\|_2^2\\ &+\sum_{i=1}^{n}\alpha_i\left(h-(I_i\Omega_I)W\mathbf{y}_i^\top-\xi_i\right)-\sum_{i=1}^{n}\beta_i\xi_i\\ &+\sum_{i=1}^{n}\gamma_i\left(h-(T_i\Omega_T)W\mathbf{y}_i^\top-\varepsilon_i\right)-\sum_{i=1}^{n}\delta_i\varepsilon_i\\ &-Tr\left(\Gamma^\top\left(\Omega_I^\top\Omega_I-I_{d\times d}\right)\right)-Tr\left(\Delta^\top\left(\Omega_T^\top\Omega_T-I_{d\times d}\right)\right),\end{aligned} \tag{6}$$

where $\alpha_i,\ \beta_i,\ \gamma_i,\ \delta_i, i=1,\cdots,n,\ \Gamma$ and Δ are Lagrange multiplier variables. To seek the minimization of the objective, we set the partial derivative of $\mathcal{L}$ with regard to W, ξ_i and ε_i to zero respectively,

$$\begin{aligned}&\frac{\partial\mathcal{L}}{\partial W} = W-\sum_{i=1}^{n}\alpha_i(I_i\Omega_I)^\top\mathbf{y}_i-\sum_{i=1}^{n}\gamma_i(T_i\Omega_T)^\top\mathbf{y}_i=0,\\ &\Rightarrow W=\sum_{i=1}^{n}\alpha_i(I_i\Omega_I)^\top\mathbf{y}_i+\sum_{i=1}^{n}\gamma_i(T_i\Omega_T)^\top\mathbf{y}_i,\\ &\frac{\partial\mathcal{L}}{\partial\xi_i}=C_1-\alpha_i-\beta_i=0\Rightarrow C_1-\alpha_i=\beta_i\geq 0\Rightarrow\alpha_i\leq C_1\\ &\frac{\partial\mathcal{L}}{\partial\varepsilon_i}=C_1-\gamma_i-\delta_i=0\Rightarrow C_1-\gamma_i=\delta_i\geq 0\Rightarrow\gamma_i\leq C_1.\end{aligned} \tag{7}$$

By substituting (7) to (6), we have

$$\mathcal{L} = f(\alpha_i|_{i=1}^n,\gamma_i|_{i=1}^n,\Omega_I,\Omega_T)+g(\alpha_i|_{i=1}^n,\gamma_i|_{i=1}^n,\Omega_I,\Omega_T,\Delta,\Gamma) \tag{8}$$

where

$$\begin{aligned}&f(\alpha_i|_{i=1}^n,\gamma_i|_{i=1}^n,\Omega_I,\Omega_T)\\ &=-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j Tr\left(\Omega_I^\top I_i^\top\mathbf{y}_i\mathbf{y}_j^\top I_j\Omega_I\right)\\ &+\sum_{i,j=1}^{n}\alpha_i\gamma_j Tr\left(\Omega_I I_i\mathbf{y}_i^\top\mathbf{y}_j T_j\Omega_T\right)\\ &-\frac{1}{2}\sum_{i,j=1}^{n}\gamma_i\gamma_j Tr\left(\Omega_T^\top T_i^\top\mathbf{y}_i\mathbf{y}_j^\top T_j\Omega_T\right)\\ &+C_2\sum_{i=1}^{n}Tr\left(\Omega_I^\top I_i^\top I_i\Omega_I\right)-2C_2\sum_{i=1}^{n}Tr\left(\Omega_I^\top I_i^\top T_i\Omega_T\right)+\\ &C_2\sum_{i=1}^{n}Tr\left(\Omega_T^\top T_i^\top T_i\Omega_T\right),\ and\\ &g(\alpha_i|_{i=1}^n,\gamma_i|_{i=1}^n,\Omega_I,\Omega_T,\Delta,\Gamma)\\ &=h\sum_{i=1}^{n}(\alpha_i+\gamma_i)\\ &-Tr\left(\Gamma^\top\left(\Omega_I^\top\Omega_I-I_{d\times d}\right)\right)-Tr\left(\Delta^\top\left(\Omega_T^\top\Omega_T-I_{d\times d}\right)\right).\end{aligned} \tag{9}$$

The optimization problem is transferred to the following coupled problem,

$$\begin{aligned}\max_{\alpha_i|_{i=1}^n,\gamma_i|_{i=1}^n,\Gamma,\Delta}\ \min_{\Omega_I,\Omega_T}\ &\mathcal{L},\\ s.t.\ 0\leq\alpha_i\leq C_1, 0\leq\gamma_i\leq C_1, i=1,\cdots,n,&\\ \Gamma\geq 0,\Delta\geq 0.&\end{aligned} \tag{10}$$

To solve this problem, we employ the alternate optimization strategy. We optimize $\alpha_i|_{i=1}^n,\gamma_i|_{i=1}^n$ and $\Gamma,\Delta,\Omega_I,\Omega_T$ alternately in an iterative algorithm. In each iteration, we first fix $\Gamma,\Delta,\Omega_I,\Omega_T$ and optimize $\alpha_i|_{i=1}^n,\gamma_i|_{i=1}^n$, and then we fix $\alpha_i|_{i=1}^n,\gamma_i|_{i=1}^n$ and optimize $\Gamma,\Delta,\Omega_I,\Omega_T$.

1) Optimizing $\alpha_i|_{i=1}^n$ and $\gamma_i|_{i=1}^n$: Fixing Γ,Δ,Ω_I and Ω_T, and removing the terms irrelevant to $\alpha_i|_{i=1}^n$ and $\gamma_i|_{i=1}^n$ from (8), we rewrite (10) as

$$\begin{aligned}\max_{\alpha_i|_{i=1}^n,\gamma_i|_{i=1}^n}\ &-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j Tr\left(\Omega_I^\top I_i^\top\mathbf{y}_i\mathbf{y}_j^\top I_j\Omega_I\right)\\ &+\sum_{i,j=1}^{n}\alpha_i\gamma_j Tr\left(\Omega_I I_i\mathbf{y}_i^\top\mathbf{y}_j T_j\Omega_T\right)\\ &-\frac{1}{2}\sum_{i,j=1}^{n}\gamma_i\gamma_j Tr\left(\Omega_T^\top T_i^\top\mathbf{y}_i\mathbf{y}_j^\top T_j\Omega_T\right)\\ &+h\sum_{i=1}^{n}(\alpha_i+\gamma_i)\\ s.t.\ 0\leq\alpha_i\leq C_1, 0\leq\gamma_i\leq C_1, i=1,\cdots,n.&\end{aligned} \tag{11}$$

This problem can be solved as a QP problem.

2) Optimizing Ω_I and Ω_T: By fixing $\alpha_i|_{i=1}^n$ and $\gamma_i|_{i=1}^n$, and removing terms irrelevant to Γ,Δ,Ω_I and Ω_T from (8), we rewrite (10) as

$$\begin{aligned}\max_{\Gamma,\Delta}\ \min_{\Omega_I,\Omega_T}\ & -\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j Tr\left(\Omega_I^\top I_i^\top \mathbf{y}_i\mathbf{y}_j^\top I_j\Omega_I\right)\\ &+\sum_{i,j=1}^{n}\alpha_i\gamma_j Tr\left(\Omega_I^\top I_i^\top \mathbf{y}_i\mathbf{y}_j^\top T_j\Omega_T\right)\\ &-\frac{1}{2}\sum_{i,j=1}^{n}\gamma_i\gamma_j Tr\left(\Omega_T^\top T_i^\top \mathbf{y}_i\mathbf{y}_j^\top T_j\Omega_T\right)\\ &+C_2\sum_{i=1}^{n}Tr\left(\Omega_I^\top I_i^\top I_i\Omega_I\right)\\ &-2C_2\sum_{i=1}^{n}Tr\left(\Omega_I^\top I_i^\top T_i\Omega_T\right)\\ &+\sum_{i=1}^{n}Tr\left(\Omega_T^\top T_i^\top T_i\Omega_T\right)\\ &-Tr\left(\Gamma^\top\left(\Omega_I^\top\Omega_I-I_{d\times d}\right)\right)\\ &-Tr\left(\Delta^\top\left(\Omega_T^\top\Omega_T-I_{d\times d}\right)\right),\\ s.t.\ &\Gamma\geq 0,\Delta\geq 0.\end{aligned}\tag{12}$$

The primal problem of this dual problem is

$$\begin{aligned}\min_{\Omega_I,\Omega_T}\ & -\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j Tr\left(\Omega_I^\top I_i^\top \mathbf{y}_i\mathbf{y}_j^\top I_j\Omega_I\right)\\ &+\sum_{i,j=1}^{n}\alpha_i\gamma_j Tr\left(\Omega_I^\top I_i^\top \mathbf{y}_i\mathbf{y}_j^\top T_j\Omega_T\right)\\ &-\frac{1}{2}\sum_{i,j=1}^{n}\gamma_i\gamma_j Tr\left(\Omega_T^\top T_i^\top \mathbf{y}_i\mathbf{y}_j^\top T_j\Omega_T\right)\\ &+C_2\sum_{i=1}^{n}Tr\left(\Omega_I^\top I_i^\top I_i\Omega_I\right)-2C_2\sum_{i=1}^{n}Tr\left(\Omega_I^\top I_i^\top T_i\Omega_T\right)\\ &+\sum_{i=1}^{n}Tr\left(\Omega_T^\top T_i^\top T_i\Omega_T\right)\\ s.t.\ &\Omega_I^\top\Omega_I=I_{d\times d},\Omega_T^\top\Omega_T=I_{d\times d}.\end{aligned}\tag{13}$$

Using the constrains $\Omega_I^\top\Omega_I = I_{d\times d}$ and $\Omega_T^\top\Omega_T - I_{d\times d}$, we can remove some constant terms from (13) and rewrite it as

$$\begin{aligned}\min_{\Omega_I,\Omega_T}\ & \sum_{i,j=1}^{n}\alpha_i\gamma_j Tr\left(\Omega_I^\top I_i^\top \mathbf{y}_i\mathbf{y}_j^\top T_j\Omega_T\right)\\ &-2C_2\sum_{i=1}^{n}Tr\left(\Omega_I^\top I_i^\top T_i\Omega_T\right)\\ &=Tr\left(\Omega_I^\top\left(\sum_{i,j=1}^{n}\alpha_i\gamma_j I_i^\top \mathbf{y}_i\mathbf{y}_j^\top T_j-2C_2\sum_{i=1}^{n}I_i^\top T_i\right)\Omega_T\right)\\ s.t.\ &\Omega_I^\top\Omega_I=I_{d\times d},\Omega_T^\top\Omega_T=I_{d\times d}.\end{aligned}\tag{14}$$

The equivalent problem is

$$\begin{aligned}\max_{\Omega_I,\Omega_T}\ & Tr\left(\Omega_I^\top Z\Omega_T\right)\\ s.t.\ &\Omega_I^\top\Omega_I=I_{d\times d},\Omega_T^\top\Omega_T=I_{d\times d}.\end{aligned}\tag{15}$$

where

$$Z=\sum_{i,j=1}^{n}\alpha_i\gamma_j I_i^\top \mathbf{y}_i\mathbf{y}_j^\top T_j-2C_2\sum_{i=1}^{n}I_i^\top T_i\in\mathbb{R}^{d_I\times d_T}.\tag{16}$$

The optimal matrices Ω_I and Ω_T can be obtained by a SVD of the matrix Z, i.e.,

$$Z=\Omega_I\Sigma\Omega_T^\top\tag{17}$$

where Σ is the matrix of singular values of Z, and Σ is diagonal.

C. Algorithm

Based on the optimization results, we can design an iterative algorithm as Algorithm 1. The iterations are repeated T times, and the t-th each iteration, the variables Ω_I^t, Ω_T^t, $\alpha_i^t|_{i=1}^n$ and $\gamma_i^t|_{i=1}^n$ are updated alternately. Finally the predictor parameter W is calculated from the updated variables.

Algorithm 1 Iterative learning algorithm of supervised cross-modal factor analysis.

Input: A training set of n documents $\{(I_1,T_i),\cdots,(I_n,T_n)\}$, and corresponding class label vector set $\{\mathbf{y}_1,\cdots,\mathbf{y}_n\}$;
Input: Tradeoff parameters C_1, C_2, and maximum iteration number T;
Initialize orthonormal transformation matrices Ω_I^0 and Ω_T^0;
for $t=1,\cdots,T$ **do**
  Fix Ω_I^{t-1} and Ω_T^{t-1}, and update $\alpha_i^t|_{i=1}^n$ and $\gamma_i^t|_{i=1}^n$ by solving the problem in (11);
  Fix $\alpha_i^t|_{i=1}^n$ and $\gamma_i^t|_{i=1}^n$, and update Z^t as in (16);
  Update Ω_I^t and Ω_T^t by applying SVD to Z^t as in (17);
end for
Calculate predictor function parameter matrix W from Ω_I^T, Ω_T^T, $\alpha_i^T|_{i=1}^n$ and $\gamma_i^T|_{i=1}^n$ (7).
Output: The orthonormal transformation matrices Ω_I^T and Ω_T^T and predictor function parameter matrix W.

III. Experiments

In this section, we will investigate the proposed algorithm experimentally.

A. Experiment setup

1) Data sets: In this experiment, we used two different document data sets composed of images and texts. The first data set is the TVGraz database [17], which is a multimodal database of object categories composed of textual and visual features. The documents belongs to 10 of 256 classes of Caltech-256. 1,000 webpages are retrieved for each of the 10 classes, and 2,058 image-text pairs are collected. Each

image-text pair is linked to a document, thus we have 2,058 documents of 10 classes in this data set.

The seconde data set is the Wikipedia database, which is selected from Wikipedia featured article database, and Wikipedia featured article database contains documents of 30 classes. Because most of the classes of Wikipedia featured articles database contains very few documents, we only choose the 10 classes with the most documents. Moreover, each featured article usually have more than one image and section, so we split each featured article to several documents. Each document contains a section of a featured article, and the images placed to this section. We have in total 7,114 documents. Moreover, we remove the documents which have more than one image, and the documents which have a text with less than 70 words. Finally, we have 2,866 documents in total in our experiment.

The images in the documents are represented as feature vectors using the bag-of-features method, while the texts in the documents are represented as feature vectors using the bag-of-words method.

2) Experiment protocol: To conduct the experiment, we used the 10-fold cross validation strategy. The entire data set is split to 10 folds randomly, and each fold was used as a test set in turn, while the remaining 9 sets were combined and used as a training set. The proposed learning algorithm was performed to the training set to learn the orthonormal transformation matrices and the predictor parameter matrix, and then they are used to represent and classify the individual images and texts in the test set. The classification accuracy is measured by the classification rate as follows,

$$\begin{aligned} &classification\ rate \\ &= \frac{number\ of\ correctly\ classified\ images\ and\ texts}{total\ number\ of\ test\ images\ and\ texts}. \end{aligned} \tag{18}$$

B. Results

1) Comparison to unsupervised CFA: We first compare the proposed supervised version of CFA against unsupervised CFA methods on the problem of image/text classification problem. We considered the original CFA method [11] and its kernel version (CFA$_{ker}$) [12] as data representation methods, and used a SVM as a classifier. The boxplots of classification rates of the 10-fold cross validations of the compared methods over two data sets are given in Fig. 1. It is clear that the proposed SupCFA outperforms the two unsupervised CFA methods completely. The low quartile of the SupCFA classification rates is even higher than the upper quartiles of the tow compared methods. This is not surprising at all because SupCFA is the only method which can explore the supervision information to improve the discriminative ability of cross-modal factor analysis, while CFA and CFA$_{ker}$ ignore the class label information at all. Moreover, it seems that CFA$_{ker}$ outperforms CFA due to its usage of kernel tricks.

2) Algorithm convergency: Since the proposed algorithm ia an iterative algorithm, it is important to study its convergency over iterations. We plot the curve of objective function over different iterations in Fig. 2. It could be seen that the objective function is reduced significantly in the first 30 iterations, and it tends to converge after the 30-th iteration.

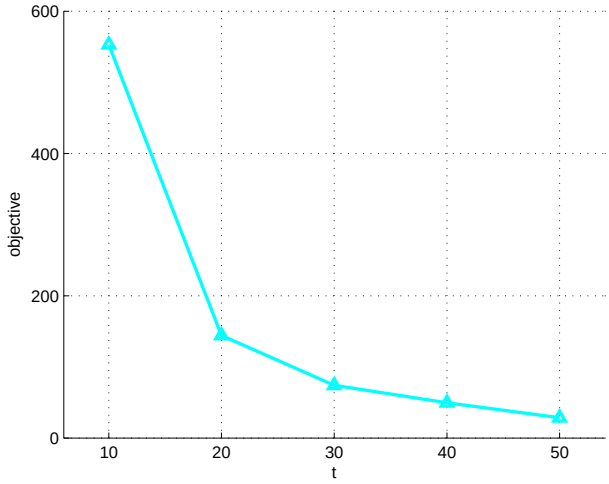


Fig. 2. Convergency curve.

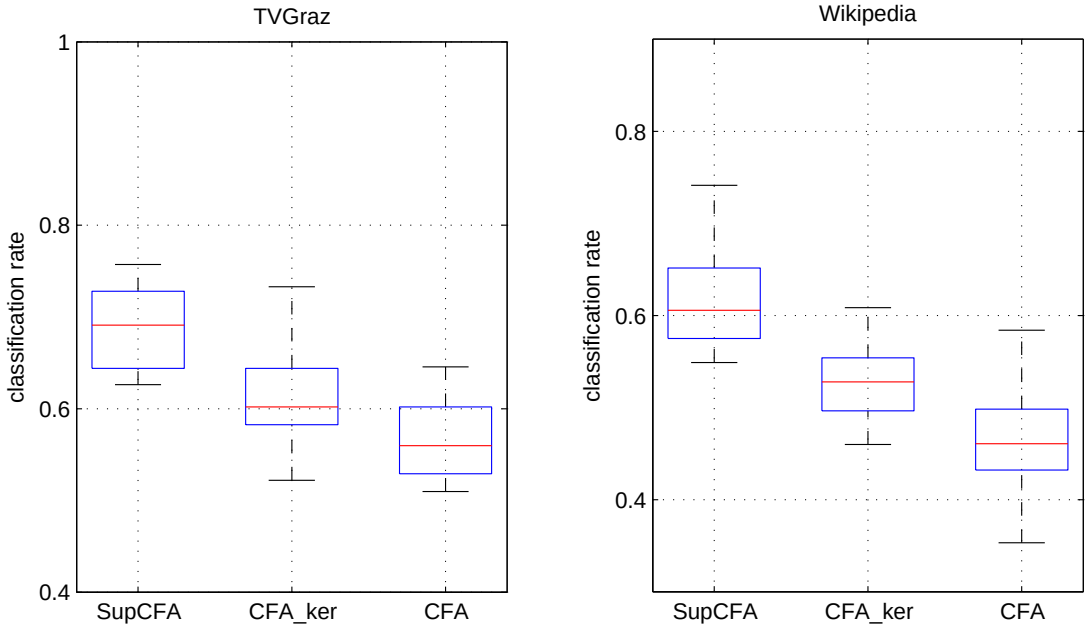
IV. Conclusions and Future works

In this paper, we proposed the first supervised CFA method for the presentation and classification of multiple modal data. The proposed method not only projects data of different modals to a shared data space like CFA, but also tries to learn a predictor to predict the class labels from this data space. Moreover, the learning of projection and class label prediction parameters are learned within a single objective function. By optimizing this objective function with regard to both projection and class label prediction parameters jointly, the class label information is used to guild the learning of CFA parameters. The experiment results show that the supervised CFA outperforms both linear and kernel versions of CFA without considering class label information. Our method can also be used in other applications such as big data analysis using high performance computing [18], [19], [20], [21], [22], [18], [23], image representation [24], information and network security [25], [26], [27], [28], [29], [30], [31], [32], and bioinformatics [33], [34], [35], [33], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [38], [39]. In this paper, we extend factor analysis for supervised cross-modal learning problem. In the future, we will also investigate the usage of some other data representation methods, such as non-negative matrix factorization [46], [47], [48], feature selection [49], [50], sparse coding [51], [52], [53], bag-of-features [54], and so on.

References

[1] F. Carenzi, P. Bendahan, V. Roschin, A. Frolov, P. Gorce, and M. Maier, "A generic neural network for multi-modal sensorimotor learning," *Neurocomputing*, vol. 58-60, pp. 525–533, 2004.

[2] A. Lumini and L. Nanni, "An advanced multi-modal method for human authentication featuring biometrics data and tokenised random numbers," *Neurocomputing*, vol. 69, no. 13-15, pp. 1706–1710, 2006.



(a) TVGraz

(b) Wikipedia

Fig. 1. Boxplots of classification rates of 10-fold cross validation of different CFA methods.

[3] K.-S. Lee, A. Nurzid Rosli, I. Ariesthea Supandi, and G.-S. Jo, "Dynamic sampling-based interpolation algorithm for representation of clickable moving object in collaborative video annotation," *Neurocomputing*, vol. 146, pp. 291–300, 2014.

[4] P. Szymczyk and M. Szymczyk, "Classification of geological structure using ground penetrating radar and laplace transform artificial neural networks," *Neurocomputing*, vol. 148, pp. 354–362, 2015.

[5] D. Merkl, "Text classification with self-organizing maps: Some lessons learned," *Neurocomputing*, vol. 21, no. 1-3, pp. 61–77, 1998.

[6] H.-J. Kim, J.-U. Kim, and Y.-G. Ra, "Boosting naïve bayes text classification using uncertainty-based selective sampling," *Neurocomputing*, vol. 67, no. 1-4 SUPPL., pp. 403–410, 2005.

[7] J. Caicedo, J. BenAbdallah, F. González, and O. Nasraoui, "Multimodal representation, indexing, automated annotation and retrieval of image collections via non-negative matrix factorization," *Neurocomputing*, vol. 76, no. 1, pp. 50–60, 2012.

[8] C. Hong and J. Zhu, "Hypergraph-based multi-example ranking with sparse representation for transductive learning image retrieval," *Neurocomputing*, vol. 101, pp. 94–103, 2013.

[9] J. Costa Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. Lanckriet, R. Levy, and N. Vasconcelos, "On the role of correlation and abstraction in cross-modal multimedia retrieval," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 3, pp. 521–535, March 2014.

[10] Y. Chen, L. Wang, W. Wang, and Z. Zhang, "Continuum regression for cross-modal multimedia retrieval," in *2012 19th IEEE International Conference on Image Processing (ICIP 2012)*, 2012, pp. 1949 – 52.

[11] D. Li, N. Dimitrova, M. Li, and I. K. Sethi, "Multimedia content processing through cross-modal association," in *Proceedings of the eleventh ACM international conference on Multimedia*. ACM, 2003, pp. 604–611.

[12] Y. Wang, L. Guan, and A. N. Venetsanopoulos, "Kernel cross-modal factor analysis for multimodal information fusion," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2384–2387.

[13] F. Liu, G. Yang, Y. Yin, and S. Wang, "Singular value decomposition based minutiae matching method for finger vein recognition," *Neurocomputing*, vol. 145, pp. 75–89, 2014.

[14] X. Zhang, Z. Xu, N. Jia, W. Yang, Q. Feng, W. Chen, and Y. Feng, "Denoising of 3d magnetic resonance images by using higher-order singular value decomposition," *Medical Image Analysis*, vol. 19, no. 1, pp. 75–86, 2015.

[15] P. Miao, Y. Shen, and X. Xia, "Finite time dual neural networks with a tunable activation function for solving quadratic programming problems and its application," *Neurocomputing*, vol. 143, pp. 80–89, 2014.

[16] F. Fomeni and A. Letchford, "A dynamic programming heuristic for the quadratic knapsack problem," *INFORMS Journal on Computing*, vol. 26, no. 1, pp. 173–182, 2014.

[17] I. Khan, A. Saffari, and H. Bischof, "Tvgraz: Multi-modal learning of object categories by combining textual and visual features," in *AAPR Workshop*, 2009, pp. 213–224.

[18] F. Zhang, Y. Gao, and J. D. Bakos, "Lucas-kanade optical flow estimation on the ti c66x digital signal processor," in *High Performance Extreme Computing Conference (HPEC), 2014 IEEE*, 2014, pp. 1–6.

[19] Y. Gao, F. Zhang, and J. D. Bakos, "Sparse matrix-vector multiply on the keystone ii digital signal processor," in *High Performance Extreme Computing Conference (HPEC), 2014 IEEE*, 2014, pp. 1–6.

[20] Y. Zhang, F. Zhang, Z. Jin, and J. D. Bakos, "An fpga-based accelerator for frequent itemset mining," *ACM Transactions on Reconfigurable Technology and Systems (TRETS)*, vol. 6, no. 1, p. 2, 2013.

[21] Y. Zhang, F. Zhang, and J. Bakos, "Frequent itemset mining on large-scale shared memory machines," in *Cluster Computing (CLUSTER), 2011 IEEE International Conference on*, 2011, pp. 585–589.

[22] F. Zhang, Y. Zhang, and J. D. Bakos, "Accelerating frequent itemset mining on graphics processing units," *The Journal of Supercomputing*, vol. 66, no. 1, pp. 94–117, 2013.

[23] F. Zhang, Y. Zhang, and J. Bakos, "Gpapriori: Gpu-accelerated frequent itemset mining," in *Cluster Computing (CLUSTER), 2011 IEEE International Conference on*, 2011, pp. 590–594.

[24] H. Wang and J. Wang, "An effective image representation method using kernel classification," in *Tools with Artificial Intelligence (ICTAI), 2014 IEEE 26th International Conference on*, Nov 2014, pp. 853–858.

[25] L. Xu, Z. Zhan, S. Xu, and K. Ye, "An evasion and counter-evasion study in malicious websites detection," in *Communications and Network Security (CNS), 2014 IEEE Conference on*, 2014, pp. 265–273.

[26] Z. Zhan, M. Xu, and S. Xu, "Characterizing honeypot-captured cyber attacks: Statistical framework and case study," *Information Forensics and Security, IEEE Transactions on*, vol. 8, no. 11, pp. 1775–1789, 2013.

[27] L. Xu, Z. Zhan, S. Xu, and K. Ye, "Cross-layer detection of malicious websites," in *Proceedings of the third ACM conference on Data and application security and privacy*. ACM, 2013, pp. 141–152.

[28] Q. Sun, P. Wu, Y. Wu, M. Guo, and J. Lu, "Unsupervised multi-level non-negative matrix factorization model: Binary data case," *Journal of Information Security*, vol. 3, no. 04, p. 245, 2012.

[29] Q. Sun, F. Hu, Y. Wu, and X. Huang, "Primate-inspired adaptive routing in intermittently connected mobile communication systems," *Wireless Networks*, vol. 20, no. 7, pp. 1939–1954, 2014.

[30] Q. Sun, W. Yu, N. Kochurov, Q. Hao, and F. Hu, "A multi-agent-based intelligent sensor and actuator network design for smart house and home automation," *Journal of Sensor and Actuator Networks*, vol. 2, no. 3, pp. 557–588, 2013.

[31] Q. Sun, F. Hu, and Q. Hao, "Human movement modeling and activity perception based on fiber-optic sensing system," *Human-Machine Systems, IEEE Transactions on*, vol. 44, no. 6, pp. 743–754, 2014.

[32] ——, "Mobile target scenario recognition via low-cost pyroelectric sensing system: Toward a context-enhanced accurate identification," *Systems, Man, and Cybernetics: Systems, IEEE Transactions on*, vol. 44, no. 3, pp. 375–384, 2014.

[33] J. Hu and F. Zhang, "Improving protein localization prediction using amino acid group based physichemical encoding," in *Bioinformatics and Computational Biology*, 2009, pp. 248–258.

[34] F. Zhang and J. Hu, "Bioinformatics analysis of physicochemical properties of protein sorting signals," 2010.

[35] ——, "Bayesian classifier for anchored protein sorting discovery," in *Bioinformatics and Biomedicine, 2009. BIBM'09. IEEE International Conference on*, 2009, pp. 424–428.

[36] M. Bhuyan and X. Gao, "A protein-dependent side-chain rotamer library." *BMC bioinformatics*, vol. 12 Suppl 14, p. S10, 2011.

[37] J. J.-Y. Wang, H. Bensmail, and X. Gao, "Multiple graph regularized protein domain ranking," *BMC bioinformatics*, vol. 13, no. 1, p. 307, 2012.

[38] Y. Liu, J. Yang, Y. Zhou, and J. Hu, "Structure design of vascular stents," *Multiscale simulations and mechanics of biological materials*, pp. 301–317, 2013.

[39] Y. Zhou, W. Hu, B. Peng, and Y. Liu, "Biomarker binding on an antibody-functionalized biosensor surface: the influence of surface properties, electric field, and coating density," *The Journal of Physical Chemistry C*, vol. 118, no. 26, pp. 14 586–14 594, 2014.

[40] J. J.-Y. Wang, X. Wang, and X. Gao, "Non-negative matrix factorization by maximizing correntropy for cancer clustering," *BMC bioinformatics*, vol. 14, no. 1, p. 107, 2013.

[41] J. Wang, Y. Li, Q. Wang, X. You, J. Man, C. Wang, and X. Gao, "Proclusensem: predicting membrane protein types by fusing different modes of pseudo amino acid composition," *Computers in biology and medicine*, vol. 42, no. 5, pp. 564–574, 2012.

[42] J. Wang, X. Gao, Q. Wang, and Y. Li, "Prodis-contshc: learning protein dissimilarity measures and hierarchical context coherently for protein-protein comparison in protein database retrieval," *BMC bioinformatics*, vol. 13, no. Suppl 7, p. S2, 2012.

[43] L. Dai, X. Gao, Y. Guo, J. Xiao, Z. Zhang *et al.*, "Bioinformatics clouds for big data manipulation," *Biology direct*, vol. 7, no. 1, p. 43, 2012.

[44] B. Peng, Y. Liu, Y. Zhou, L. Yang, G. Zhang, and Y. Liu, "Modeling nanoparticle targeting to a vascular surface in shear flow through diffusive particle dynamics," *Nanoscale Research Letters*, vol. 10, no. 1, p. 235, 2015.

[45] S. Wang, Y. Zhou, J. Tan, J. Xu, J. Yang, and Y. Liu, "Computational modeling of magnetic nanoparticle targeting to stent surface under high gradient field," *Computational mechanics*, vol. 53, no. 3, pp. 403–412, 2014.

[46] J. J.-Y. Wang, H. Bensmail, and X. Gao, "Multiple graph regularized nonnegative matrix factorization," *Pattern Recognition*, vol. 46, no. 10, pp. 2840–2847, 2013.

[47] J. J.-Y. Wang and X. Gao, "Beyond cross-domain learning: Multiple-domain nonnegative matrix factorization," *Engineering Applications of Artificial Intelligence*, vol. 28, pp. 181–189, 2014.

[48] Q. Sun, J. Lu, Y. Wu, H. Qiao, X. Huang, and F. Hu, "Non-informative hierarchical bayesian inference for non-negative matrix factorization," *Signal Processing*, vol. 108, pp. 309–321, 2015.

[49] J. J.-Y. Wang, H. Bensmail, and X. Gao, "Feature selection and multi-kernel learning for sparse representation on a manifold," *Neural Networks*, vol. 51, pp. 9–16, 2014.

[50] J.-Y. Wang, I. Almasri, and X. Gao, "Adaptive graph regularized nonnegative matrix factorization via feature selection," in *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 2012, pp. 963–966.

[51] J. J.-Y. Wang, H. Bensmail, N. Yao, and X. Gao, "Discriminative sparse coding on multi-manifolds," *Knowledge-Based Systems*, vol. 54, pp. 199–206, 2013.

[52] J. Luo and A. Brodsky, "Piecewise surface regression modeling in intelligent decision guidance system," in *Intelligent Decision Technologies*. Springer, 2011, pp. 223–235.

[53] ——, "Piecewise regression learning in corejava framework," *International Journal of Machine Learning and Computing*, vol. 1, no. 2, pp. 163–169, 2011.

[54] J. J.-Y. Wang, H. Bensmail, and X. Gao, "Joint learning and weighting of visual vocabulary for bag-of-feature based tissue classification," *Pattern Recognition*, vol. 46, no. 12, pp. 3249–3255, 2013.