2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Integrating Supervised Subspace Criteria with Restricted Boltzmann Machine for Feature Extraction

Guo-Sen Xie, Xu-Yao Zhang, Yan-Ming Zhang and Cheng-Lin Liu
National Laboratory of Pattern Recognition (NLPR)
Institute of Automation, Chinese Academy of Sciences
Beijing 100190, China
Email: {guosen.xie, xyz, ymzhang, liucl}@nlpr.ia.ac.cn

***Abstract*—Restricted Boltzmann Machine (RBM) is a widely used building-block in deep neural networks. However, RBM is an unsupervised model which can not exploit the rich supervised information of data. Therefore, we consider combining the descriptive (generative) ability of RBM with the discriminative ability of supervised subspace models, i.e., Fisher linear discriminant analysis (FDA), marginal Fisher analysis (MFA), and heat kernel MFA (hkMFA). Specifically, the hidden layer of RBM is regularized by the supervised subspace criteria, and the joint learning model can then be efficiently optimized by gradient descent and graph construction (used to define the scatter matrix in the subspace models) on mini-batch data. Compared with the traditional subspace models (FDA, MFA, hkMFA), the proposed hybrid models are essentially nonlinear and can be optimized by gradient descent instead of eigenvalue decomposition. More importantly, traditional subspace models can only reduce the dimensionality (because of linear transformation), while the proposed models can also increase the dimensionality for better class discrimination. Experiments on three databases demonstrate that the proposed hybrid models outperform both RBM and their counterpart subspace models (FDA, MFA, hkMFA) consistently.**

I. Introduction

Restricted Boltzmann Machine (RBM) [1] [2] is a bipartite undirected graphical model, and is also the building block of many complex generative models such as Deep Belief Networks (DBN), Deep Boltzmann Machine (DBM) and Auto-Encoders [3] [4]. RBM related models have become very popular because of their success in various application domains, such as dimensionality reduction, classification, information retrieval and so on [3]–[6].

Recently, deep neural networks (DNN) pre-trained layer-wisely have shown great success in various domains [3], [7], [8]. As one kind of building block of DNN, RBM has drawn much attention in the research community. One single layer RBM is a generative model which can model the probability density function of input samples [2]. Commonly, RBM consists of input layer, hidden layer and the connections between these two layers. There are neither connections within the input units nor within the hidden units (Fig. 1). With this connection relationship, RBM can be inferred efficiently by the gradient-based Contrastive Divergence (CD) algorithm [2].

This work was supported by National Basic Research Program of China (973 Program) Grant 2012CB316302 and National Natural Science Foundation of China (NSFC) Grant 61203296.

Many variants of RBM have been developed since Hinton et al. [2] [3] proposed the the pioneer work about CD algorithm and layer-wise pre-training. Lee et al. [9] proposed to train a RBM with sparse constraint of hidden output, which holds promise for modeling higher-order features. To add discrimination to RBM, Nair et al. [10] presented a three-order RBM model, which can model the top-level joint distribution in which the class label multiplicatively interacts with both the penultimate layer units and the output units to determine the energy of a full configuration. The hybrid inferring algorithm of three-order RBM consists of generative training followed by discriminative updating. Larochelle et al. [11] viewed RBM as a classifier (not only as feature extractor) through modeling the joint distribution of the inputs and associated targets. Stuhlsatz et al. [12] performed supervised pre-training by extending the output of RBM with extra visual output targets. Then, discriminant criterion evaluated in the hidden space can be asymptotically maximized by minimizing the mean squared error between outputs and according targets.

With the development of manifold learning, many graph based subspace learning algorithms were proposed [13]–[16]. When deep learning encounters manifold learning, some methods combining them have been presented. Wong et al. [17] proposed regularized deep Fisher mapping, which adds reconstruction (based on auto-encoder) and weight decay as its regularization of parameters. Weston et al. [18] combined an embedding-based regularizer with some (or all) layers of a deep supervised learner to perform semi-supervised learning. Salakhutdinov and Hinton [19] presented nonlinear Neighborhood Components Analysis (NCA) with auto-encoder regularizer to reconstruct the data from the coding. Yu et al. [20] developed unsupervised embedding with auto-encoder reconstruction, which can support incremental embedding because of the exact calculation of the embedding weights.

In order to obtain discriminative features, the method of [10] is based on generative learning of RBM followed by label induced discriminative learning. The learning method of [11] is based on variant CD algorithm with the assumption of joint distribution of label and data. Moreover, the training method of [12] is based on minimum squared error (MSE). None of the above methods incorporates supervised subspace criteria on the hidden layer of RBM. Inspired by recently proposed subspace learning algorithms, we consider imposing discriminative subspace constraint during training process of RBM so that the weights are updated simultaneously. Specif-

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

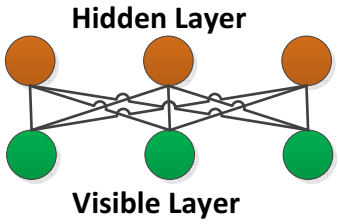

Fig. 1. The diagram of RBM.

ically, we consider three discriminative subspace regularizing methods: (a) Fisher Linear Discriminative Analysis (FDA) based on pairwise definition of within-class and between-class scatter matrices [21], (b) Marginal Fisher Analysis (MFA) [16], and (c) heat kernel form of MFA (hkMFA). Joint training algorithm of RBM with the supervised subspace (FDA, MFA, hkMFA) regularization is then optimized efficiently by gradient descent based on mini-batch data. The gradient update of the training process is composed of two parts: one is from the CD algorithm, the other from the gradient increment of subspace constraint w.r.t. the weights. Due to joint weight updating of the two parts, the model can hold both discriminative and descriptive abilities. In the learning process, the subspace constraint is incorporated by the graph construction (associated with scatter matrices in FDA, MFA, and hkMFA) based on the batch data used to calculate gradients. Therefore, the training process is as efficient as the traditional RBM model. From the viewpoint of dimension reduction (DR), the proposed hybrid models (FDA-RBM, MFA-RBM and hkMFA-RBM) can be seen as nonlinear feature extractors with data reconstruction constraint from CD algorithm. Hence, they can increase the dimensionality to extract much more powerful features by considering the generative and discriminative information. Our experimental results show that the proposed hybrid models can outperform both the non-regularized unsupervised RBM and the baseline supervised subspace models (FDA, MFA, hkMFA).

The rest of this paper is organized as follows: The preliminaries about RBM and contrastive divergence are presented in Section II. The proposed hybrid models are detailed in Section III-V. Experimental results are presented in Section VI. Finally, Section VII concludes the paper.

II. RBM and Contrastive Divergence

RBM (Fig. 1) is a two-layer connecting network. Here, we assume that both the input and hidden units of the RBM are binary, denoted by $x \in \{0,1\}^D$ and $h \in \{0,1\}^F$ respectively. The energy function of the RBM is:

$$E(x,h) = -\sum_{i=1}^{D}\sum_{j=1}^{F} x_i w_{ij} h_j - \sum_{i=1}^{D} b_i x_i - \sum_{j=1}^{F} c_j h_j, \quad (1)$$

where the parameters $\theta = \{W = [w_{ij}]^{D\times F}, b = [b_i]^D, c = [c_j]^F\}$ are learned from data. w_{ij} represents the interaction weight between input visible unit i and hidden unit j. b_i and c_j are the biases of the visible and hidden units respectively. The joint distribution of (x,h) is defined as:

$$P(x,h;\theta) = \frac{\exp(-E(x,h;\theta))}{Z(\theta)}, \quad (2)$$

where $Z(\theta) = \sum_x \sum_h \exp(-E(x,h;\theta))$. The probability assigned by the distribution to a visible vector is:

$$P(x;\theta) = \frac{\sum_h \exp(-E(x,h;\theta))}{Z(\theta)}. \quad (3)$$

The conditional probability (on hidden and visible units) are

$$p(h_j = 1 \mid x) = \sigma(\sum_{i=1}^{D} w_{ij} x_i + c_j), \quad (4)$$

$$p(x_i = 1 \mid h) = \sigma(\sum_{j=1}^{F} w_{ij} h_j + b_i), \quad (5)$$

where $\sigma(x) = \frac{1}{1+\exp(-x)}$. The generative training of RBM is based on maximizing the log likelihood: $L(\theta)$ of the joint distribution of train data. Given the training data $\{x^{(t)}, y^{(t)}\}_{t=1}^N$. $L(\theta)$ is written as: $L(\theta) = \sum_{t=1}^N \log P(x^{(t)};\theta)$. Taking partial derivative of $L(\theta)$ w.r.t. θ, we can get:

$$\frac{\partial L}{\partial \theta} = \sum_{t=1}^{N} \langle \frac{\partial(-E(x^{(t)},h))}{\partial\theta} \rangle_{P(h|x^{(t)})} - \langle \frac{\partial(-E(x,h))}{\partial\theta} \rangle_{P(x,h)}, \quad (6)$$

where $\langle\rangle_P$ is the expectation w.r.t. the distribution P. Exact computation of the expectation of the second term in (6) takes exponential time, which makes the exact maximum likelihood learning intractable for large data size.

In practice, we can get an approximation to the gradient of a different function by contrastive divergence (CD) [2]. $\langle\cdot\rangle_{P(x,h)}$ is approximated by $\langle\cdot\rangle_{P_T}$, where P_T is a distribution defined by running a Gibbs chain for T steps. Setting $T = \infty$ recovers the maximum likelihood learning of the model.

Learning in the above model is often performed based on mini-batch, wherein the summation of (6) are computed for only a small subset of K training samples (in subsequent experiments, we take mini-batch size as 100). Substitute θ with w_{ij}, b_i, c_j, the gradient increments are specified as

$$\Delta w_{ij} = \frac{1}{K}\sum_{t=1}^{K} p(h_j = 1|x^{(t)}) x_i^{(t)} - p(h_j = 1|x^{(t)-}) x_i^{(t)-}, \quad (7)$$

$$\Delta c_j = \frac{1}{K}\sum_{t=1}^{K} p(h_j = 1|x^{(t)}) - p(h_j = 1|x^{(t)-}), \quad (8)$$

$$\Delta b_i = \frac{1}{K}\sum_{t=1}^{K} x_i^{(t)} - x_i^{(t)-}, \quad (9)$$

where K is the batch-size, $x^{(t)-}$ is sampled from $p(x|h^{(t)})$ and $h^{(t)}$ is sampled from $p(h|x^{(t)})$.

III. FDA Regularized RBM

In this Section, the supervised Fisher criteria (FDA) [22] [23] is incorporated as regularization on the hidden layer of RBM to learn joint generative and discriminative model. The RBM model is trained by the CD algorithm (Section II) which is carried out based on mini-batch gradient descent. Given the batch data $\{x^{(t)}, y^{(t)}\}_{t=1}^{K}$, let $x^{(t)} \in \mathbb{R}^D (t = 1, 2, \cdots, K)$ be D-dimensional samples and $y^{(t)} \in \{1, 2, \cdots, C\}$ be the class labels, where K is the number of samples and C is the number of classes. Let N_c be the number of samples in class c: $\sum_{c=1}^{C} N_c = K$. Let $h^{(t)} = \frac{1}{1+e^{-W^\top x^{(t)} - c}} \in \mathbb{R}^F (t = 1, 2, \cdots, K)$ be the hidden outputs of the RBM. The parameters: $W \in \mathbb{R}^{D \times F}$ and $c \in \mathbb{R}^F$ should be calculated based on not only the CD Algorithm but also on the FDA like criteria on the hidden outputs of RBM. The optimization objective can be formulated as follow:

$$\begin{aligned} \max_{\theta} \mathcal{L} &= L_1 - \lambda_1 L_2 \\ &= \sum_{t=1}^{K} \log \frac{\sum_h \exp(-E(x^{(t)}, h; \theta))}{Z(\theta)} - \lambda_1 \frac{tr(S^{(w)})}{tr(S^{(b)})}. \end{aligned} \tag{10}$$

The hyper-parameter λ_1 control the balance between the generative objective of RBM L_1 and discriminative objective of FDA L_2. The within-class scatter matrix $S^{(w)}$ and between-class scatter matrix $S^{(b)}$ are defined according to the FDA criterion [21] on the hidden layer of the RBM model:

$$S^{(w)} = \frac{1}{2} \sum_{i=1}^{K} \sum_{j=1}^{K} A_{i,j}^{(w)} (h^{(i)} - h^{(j)})(h^{(i)} - h^{(j)})^\top, \tag{11}$$

$$S^{(b)} = \frac{1}{2} \sum_{i=1}^{K} \sum_{j=1}^{K} A_{i,j}^{(b)} (h^{(i)} - h^{(j)})(h^{(i)} - h^{(j)})^\top, \tag{12}$$

where

$$A_{i,j}^{(w)} = \begin{cases} \frac{1}{N_c}, & \text{if } y^{(i)} = y^{(j)} = c \\ 0, & \text{if } y^{(i)} \neq y^{(j)}, \end{cases} \tag{13}$$

$$A_{i,j}^{(b)} = \begin{cases} \frac{1}{K} - \frac{1}{N_c}, & \text{if } y^{(i)} = y^{(j)} = c \\ \frac{1}{K}, & \text{if } y^{(i)} \neq y^{(j)}. \end{cases} \tag{14}$$

$A_{i,j}^{(w)}$ and $A_{i,j}^{(b)}$ are the within-class and between-class adjacency matrix (AM) of the samples of one mini-batch data. Based on [17], the second term L_2 of (10) can be rewritten as:

$$\begin{aligned} L_2 = \frac{tr(S^{(w)})}{tr(S^{(b)})} &= \frac{\frac{1}{2} \sum_{i=1}^{K} \sum_{j=1}^{K} A_{ij}^{(w)} \parallel h^{(i)} - h^{(j)} \parallel^2}{\frac{1}{2} \sum_{i=1}^{K} \sum_{j=1}^{K} A_{ij}^{(b)} \parallel h^{(i)} - h^{(j)} \parallel^2} \\ &= \frac{\mathbf{1}_K^\top (A^{(w)} \odot O) \mathbf{1}_K}{\mathbf{1}_K^\top (A^{(b)} \odot O) \mathbf{1}_K}, \end{aligned} \tag{15}$$

where O satisfies $O_{ij} = \parallel h^{(i)} - h^{(j)} \parallel^2$ and $\mathbf{1}_K \in \mathbb{R}^K$ is a vector with all the elements equal to one. Therein $A^{(w)} = [A_{i,j}^{(w)}]^{K \times K}$, $A^{(b)} = [A_{i,j}^{(b)}]^{K \times K}$. $\odot$ indicates element-wise products between matrices. Note here the FDA criteria L_2 is defined on the hidden layer h not original space x. In this way, the supervised information can be incorporated into the learning process of the RBM model. In the mini-batch based gradient descent learning process, the gradient of L_1 can be calculated based on the CD algorithm as (7) - (9) in Section II. Meanwhile, the gradient of L_2 w.r.t. the model parameters W, c can be calculated in Algorithm 1. In Algorithm 1, the outputs in hidden layer of batch data are calculated in Step 2. For the K samples in a mini-batch, the gradients of trace of within-class and between-class scatter matrices w.r.t. hidden layers are obtained in Steps 5 and 6, respectively. After that, Steps 7-10 give the gradient of L_2 w.r.t. the model parameters. The gradients of L_1 and L_2 are then combined and the gradient descent learning is implemented based on both parts. Given the learning rate η, the parameters are updated as:

$$\begin{aligned} W &= W + \eta(\triangle W_{L_1} - \lambda_1 \triangle W_{L_2}), \\ c &= c + \eta(\triangle c_{L_1} - \lambda_1 \triangle c_{L_2}), \\ b &= b + \eta(\triangle b_{L_1}), \end{aligned} \tag{16}$$

Algorithm 1 Gradient Calculation of FDA

Input:
Batch Data: $\Phi = \{(x^{(t)}, y^{(t)}) | x^{(t)} \in \mathbb{R}^D\}$ $, t = 1, 2, \cdots, K$

Output: $\triangle W_{L_2}, \triangle c_{L_2}$

1: Initialize $\triangle W_{L_2} = 0, \triangle c_{L_2} = 0$.
2: Substitute $\{h^{(t)} | h^{(t)} = \frac{1}{1+e^{(-W^\top x^{(t)} - c)}} \in \mathbb{R}^F\}$ $, t = 1, 2, \cdots, K$ to equation (15).
3: Set $H = [h^{(1)}, h^{(2)}, \cdots, h^{(K)}] \in \mathbb{R}^{F \times K}$.
4: **for** t = 1 → K **do**
5: $\frac{\partial tr(S^{(w)})}{\partial h^{(t)}} = (h^{(t)} \mathbf{1}_K^\top - H)(A^{(w)} + A^{(w)^\top})_{(:,t)}$
6: $\frac{\partial tr(S^{(b)})}{\partial h^{(t)}} = (h^{(t)} \mathbf{1}_K^\top - H)(A^{(b)} + A^{(b)^\top})_{(:,t)}$
7: $\frac{\partial L_2}{\partial h^{(t)}} = \frac{tr(S^{(b)}) \cdot \frac{\partial tr(S^{(w)})}{\partial h^{(t)}} - tr(S^{(w)}) \cdot \frac{\partial tr(S^{(b)})}{\partial h^{(t)}}}{[tr(S^{(b)})]^2}$
8: $\delta = \frac{\partial L_2}{\partial h^{(t)}} \odot h^{(t)} \odot (\mathbf{1}_F - h^{(t)})$
9: $\triangle W_{L_2} \leftarrow \triangle W_{L_2} + \frac{1}{K} x^{(t)} * \delta^\top$
10: $\triangle c_{L_2} \leftarrow \triangle c_{L_2} + \frac{1}{K} \delta$
11: **end for**

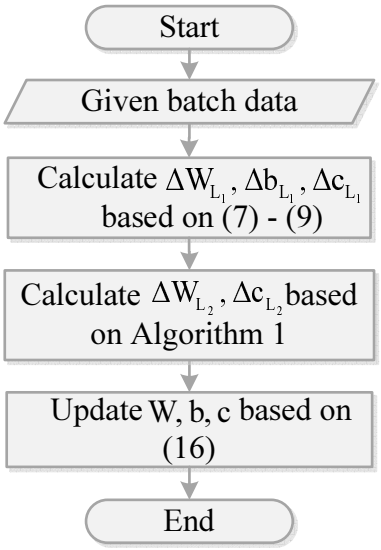

Fig. 2. The flowchart of one mini-batch based updating for the FDA regularized RBM model.

where $\triangle W_{L_1}, \triangle c_{L_1}, \triangle b_{L_1}$ are calculated based on equation (7) - (9).

To clearly depict our proposed hybrid model, we list the procedures for the updating of the model parameters on one mini-batch data in Fig 2. The mini-batch updating should run multiple times to cover all the training data until the convergence of the training process.

IV. MFA Regularized RBM

In this Section, we consider incorporating MFA [16] criteria as regularization for RBM training. For simplicity, we use the notations denoted in Section III in this part and the sequent part. We first review the concept of MFA which is proposed by [16]. MFA is based on graph embedding. An intrinsic graph characterizing the within-class compactness and a penalty graph characterizing the between-class separability are constructed respectively. The intrinsic graph on $\{x^{(t)}, y^{(t)}\}_{t=1}^{K}$ presents the intraclass adjacency relationship (each sample are connected to its k_1-nearest neighbors of the same labels). Meanwhile, the penalty graph shows the interclass point adjacency relationship (the marginal node pair of different classes are connected) (see Fig. 3 for intuition). The objective function of RBM with MFA criteria constraint on hidden layer is:

$$\begin{aligned}\max_{\theta} \mathcal{L} &= L_1 - \lambda_2 L_2 \\ &= \sum_{t=1}^{K} \log \frac{\sum_h \exp(-E(x^{(t)}, h; \theta))}{Z(\theta)} - \\ &\lambda_2 (\sum_{i=1}^{K}\sum_{j=1}^{K} (A_{ij}^{(w)} - A_{ij}^{(b)}) \parallel h^{(i)} - h^{(j)} \parallel^2).\end{aligned} \tag{17}$$

The hyper-parameter λ_2 control the balance between the generative objective of RBM L_1 and discriminative objective of MFA L_2. And $h^{(t)} = \frac{1}{1+e^{-W^\top x^{(t)} - c}} \in \mathbb{R}^F (t = 1, 2, \cdots, K)$ is the hidden layer output of the RBM model. The elements ($A_{ij}^{(w)}$ and $A_{ij}^{(b)}$) of within-class and between-class adjacency matrices for MFA are defined as [16]:

$$A_{i,j}^{(w)} = \begin{cases} 1, & \text{if } i \in \zeta_{k_1}(j) \text{ or } j \in \zeta_{k_1}(i) \\ 0, & \text{else} \end{cases} \tag{18}$$

$$A_{i,j}^{(b)} = \begin{cases} 1, & \text{if } (i,j) \in \tau_{k_2}(C_i) \text{ or } (i,j) \in \tau_{k_2}(C_j) \\ 0, & \text{else} \end{cases} \tag{19}$$

where $\zeta_{k_1}(i)$ indicates the index set of the k_1 nearest neighbor of $x^{(i)}$ in the same class as $x^{(i)}$, and $\tau_{k_2}(C_i)$ is a set of index pairs that are the the k_2 nearest pairs between class C_i and other classes.

Note that FDA adopted a quotient based criterion (15) while MFA adopted a subtraction based criterion (17), and the intra-class and inter-class adjacency weight matrices are defined differently for FDA and MFA. Therefore, they can extract different discriminative features for classification. In the experimental sections, we will compare the performance of both FDA and MFA regularized RBM models.

Similar to Algorithm 1, in the learning process, we should calculate the gradients w.r.t. the MFA criterion. The complete procedure of gradients calculation for MFA is listed in Algorithm 2. In Algorithm 2, the outputs in hidden layer of batch data are calculated in Step 2. The intra-class and inter-class adjacent matrices ($A^{(w)}$ and $A^{(b)}$) of K batch samples are constructed in Step 4. The gradient of L_2 w.r.t. hidden layer is obtained in Steps 6. Finally, Steps 7-9 give the gradient of L_2 w.r.t. the model parameters.

Algorithm 2 Gradient Calculation of MFA

Input:
Batch Data: $\Phi = \{(x^{(t)}, y^{(t)}) | x^{(t)} \in \mathbb{R}^D\}, t = 1, 2, \cdots, K$
Hyper-parameter: $k1, k2$
Output: $\triangle W_{L_2}, \triangle c_{L_2}$

1: Initialize $\triangle W_{L_2} = 0, \triangle c_{L_2} = 0$.
2: Substitute $\{h^{(t)} | h^{(t)} = \frac{1}{1+e^{(-W^\top x^{(t)} - c)}} \in \mathbb{R}^F\}, t = 1, 2, \cdots, K$ to L_2 in equation (17).
3: Set $H = [h^{(1)}, h^{(2)}, \cdots, h^{(K)}] \in \mathbb{R}^{F \times K}$.
4: Construct k_1-nearest neighbor intra-class graph: $A^{(w)}$ and k_2-nearest neighbor inter-class graph: $A^{(b)}$
5: **for** t = 1 → K **do**
6: $\frac{\partial L_2}{\partial h^{(t)}} = (2h^{(t)}\mathbf{1}_K^\top - H)((A^{(w)} - A^{(b)}) + (A^{(w)} - A^{(b)})^\top)_{(:,t)}$
7: $\delta = \frac{\partial L_2}{\partial h^{(t)}} \odot h^{(t)} \odot (\mathbf{1}_F - h^{(t)})$
8: $\triangle W_{L_2} \leftarrow \triangle W_{L_2} + \frac{1}{K} x^{(t)} \delta^\top$
9: $\triangle c_{L_2} \leftarrow \triangle c_{L_2} + \frac{1}{K}\delta$
10: **end for**

The training process of MFA regularized RBM model has the same flowchart as Fig. 2 except that the gradient $\triangle W_{L_2}, \triangle c_{L_2}$ are obtained according to Algorithm 2.

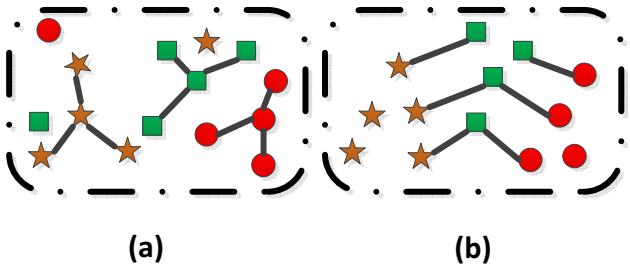

Fig. 3. The adjacency relationships of (a) Intra-class graphs and (b) Inter-class graph for the MFA

To further improve the performance, we also consider using the heat kernel method to define the adjacency matrixes [21] [23]. To reduce the number of free parameters, we set τ as the average of all Euclidean distance between nearest-neighbor samples.

$$A_{i,j}^{(w)} = \begin{cases} e^{-\frac{\|x^{(i)} - x^{(j)}\|^2}{\tau}}, & \text{if } i \in \zeta_{k_1}(j) \text{ or } j \in \zeta_{k_1}(i) \\ 0, & \text{else} \end{cases} \tag{20}$$

$$A_{i,j}^{(b)} = \begin{cases} e^{-\frac{\|x^{(i)} - x^{(j)}\|^2}{\tau}}, & \text{if } (i,j) \in \tau_{k_2}(C_i) \text{ or } \tau_{k_2}(C_j) \\ 0, & \text{else} \end{cases} \tag{21}$$

Using (20) (21) to replace (18) (19) and incorporating them into the RBM learning process (17), we can obtain the hkMFA-

RBM model which should be more effective than the MFA-RBM model.

V. Analysis of the Proposed Hybrid Models

From the viewpoint of dimensionality reduction (DR), the proposed hybrid models can be viewed as extensions of the traditional subspace models (FDA, MFA, hkMFA). Actually, the hybrid models can be seen as feature extractors with data reconstruction constraint from CD algorithm [2]. Furthermore, the new optimization problem which utilizes the two layer architecture of RBM is solved by gradient descent instead of eigenvalue decomposition. Because of the sigmoid activation in the hidden output, the proposed new models are essentially nonlinear models, however, the mapping matrix $W \in \mathbb{R}^{D \times F}$ is explicitly calculated so that the proposed algorithms can also avoid the out-of-sample extension problem [33] which is very common in other manifold based nonlinear models. Moreover, if we make the number of hidden units more than the number of input dimensionality, the new hybrid model can also increase the dimensionality. On the contrary, due to the linear transformation and eigenvalue decomposition algorithm, the traditional subspace models (FDA, MFA, hkMFA) can only reduce the dimensionality. Compared with the traditional RBM model, the proposed hybrid models can exploit much more discriminative information of the data by integrating supervised subspace criteria on the hidden layer of the RBM model, and hence, much more powerful features can be extracted.

VI. Experiments

In this Section, we compare the proposed three hybrid models (FDA-RBM, MFA-RBM, hkMFA-RBM) with the traditional RBM model, PCA model and the supervised subspace models (FDA, MFA, hkMFA). Treating each model as one feature extractor, the nearest neighbor (NN) classifier is used to evaluate the classification accuracy. All the models are evaluated on three data sets: MNIST subset [25], Pendigits [26] and the first 20 classes of Caltech101 Silhouettes [27] [28] (Fig. 4).

As for comparison, we evaluate 8 different methods including the proposed three regularized methods. These methods are

- (A) Nearest Neighbor on pixel (NN);
- (B) Principal Components Analysis (PCA);
- (C) Fisher Linear Discriminative Analysis (FDA);
- (D) Marginal Fisher Analysis (MFA);
- (E) Restricted Boltzmann Machine (RBM);
- (F) FDA Regularized RBM (FDA-RBM);
- (G) MFA Regularized RBM (MFA-RBM);
- (H) hkMFA Regularized RBM (hkMFA-RBM).

In the subsequent Sections, in order to use FDA to reduce the dimensionality to more than $C-1$ (C is class number), we use method introduced in [29] [30] to solve the FDA problem. Thus, new parameter reg (used to guarantee nonsingular of total variance in the generalized eigenvector problem) is introduced. Note that different results of FDA may be obtained, while reducing the dimension to more than

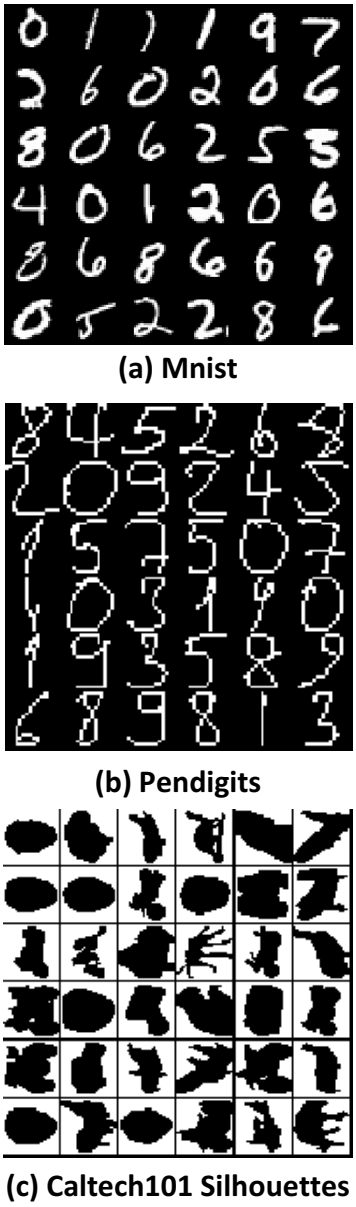

Fig. 4. Some examples of (a) MNIST (b) Pendigits and (c) Caltech101 Silhouettes.

$C-1$ on different computers. It results from the fact that you can get different eigenvectors w.r.t. the zero eigenvalue except the $C-1$ non-zero eigenvalue when doing eigenvalue decomposition on different computers. The implementation of MFA is also based on the solver introduced in [31]. To be convenient for the description, we summarize the parameters used in (A) - (H) correspondingly:

- (A) NN: no parameters;
- (B) PCA: no parameters;
- (C) FDA: reg;
- (D) MFA: k_1 (number of intra-class nearest neighbor), k_2 (number of inter-class nearest neighbor), reg;
- (E) RBM: max epoch (me), learning rate (lr), weight

decay (wd), momentum (m), steps of running Gibbs chain in Contrastive Divergence (cd), batch size (bs);

- (F) FDA-RBM: parameters of RBM, λ_1, discrimination insert position (dip), i.e., the epoch position that we start procedure in Fig. 2;
- (G) MFA-RBM: parameters of RBM, λ_2, k_1, k_2, discrimination insert position (dip);
- (H) hkMFA-RBM: parameters of MFA-RBM.

During our experiments, we found that we need not to start joint learning of RBM and supervised subspace criteria (procedure in Fig. 2) from the first iterative epoch. Actually, we only need to start jointly updating weights in the last several epoches. The rationalities of the above finding are: (1) the weights updating in the initial epoches is used as preliminary feature learning, which may be disturbed by incorporating the discriminative information; (2) after several epoches, the discriminative information added into the well-trained weights will improve the discriminative ability of features. Thus, we can obtain the regularized RBM models which must be better than the original RBM and counterpart DR methods. The details of parameter selection of (C) - (H) can be further found in Subsection A.

A. Parameter Selection

The searching space of parameters for model (C) - (H) corresponding to each data set are:

- (C) FDA: $reg \in [0, 500] \subset \mathbb{Z}$ for three data sets;
- (D) MFA: $k_1 \in [1, 100] \subset \mathbb{Z}$, $k_2 \in [0, 1000] \subset \mathbb{Z}$, $reg \in [0, 20000] \subset \mathbb{Z}$ for three data sets; Here, left interval endpoint: 0 of k_2 means full connecting of inter-class graph;
- (E) RBM: $me \in [0, 200] \subset \mathbb{Z}, cd = 3, bs = 100$ for three data sets; $lr = 0.05, wd = 0.0005, m = 0.5$ (the first 5 epoches), $m = 0.9$ (the rest epoches) for MNIST; $lr = 0.06, wd = 0.002, m = 0.4$ (the first 5 epoches), $m = 0.95$ (the rest epoches) for Pendigits; $lr = 0.03, wd = 0.002, m = 0.4$ (the first 5 epoches), $m = 0.95$ (the rest epoches) for Caltech20 Silhouettes; Mainly refer [3] [32];
- (F) FDA-RBM: parameter space of corresponding RBM for each data set; $\lambda_1 \in [1, 500] \subset \mathbb{Z}$, $dip \in [\frac{me}{2}, me] \subset \mathbb{Z}$ for three data sets;
- (G) MFA-RBM: parameter space of corresponding RBM for each data set; $10 \times \lambda_2 \in [1, 50] \subset \mathbb{Z}$, $dip \in [\frac{me}{2}, me] \subset \mathbb{Z}$, $k_1 \in [1, 100] \subset \mathbb{Z}$, $k_2 \in [1, 3000] \subset \mathbb{Z}$ for three dataset;
- (H) hkMFA-RBM: parameter space of corresponding MFA-RBM for each data set.

NN and PCA has no parameters to select. For fair comparison, we select the parameters of FDA, MFA and RBM by minimizing the test error directly on the test set. For our methods (FDA-RBM, MFA-RBM and hkMFA-RBM) on the three data sets, the following strategy is adopted to do parameter selection:

1) Fix me, lr, wd, m, cd, bs. We take these parameters the same as its corresponding RBM experiments.
2) Use five-fold cross-validation on train set to find best $\theta_1 = \{dip, \lambda_1\}$ for FDA-RBM and $\theta_2 = \{dip, \lambda_2, k_1, k_2\}$ for MFA-RBM or hkMFA-RBM. Because the searching space is very large for θ_1, θ_2. For FDA-RBM, we only search the grid of $\frac{me}{2} \times 11$ which divides interval $[\frac{me}{2}, me]$ and $[1, 500]$ respectively. Here, the 11 grid endpoints of $\lambda_1 \in [1, 500]$ are $\{1\} \cup \{x | x = 50i, i = 1, 2, \cdots, 10\}$. For MFA-RBM or hkMFA-RBM, we reduce the searching space by fixing λ_2, k_1 based on cross-validation. And, the searching grid becomes $\frac{me}{2} \times 31$ which divides interval $[\frac{me}{2}, me]$ and $[1, 3000]$ respectively. And, the 31 grid endpoints of $k_2 \in [1, 3000]$ are $\{1\} \cup \{x | x = 100i, i = 1, 2, \cdots, 30\}$. What's more, we do not carry out cross-validation on each dimensionality, parameters on one dimension are usually good enough for the other dimensions.
3) Test and record the test results based on the selected parameters.

B. MNIST Experiment

The MNIST dataset of handwritten digits (10 classes) [25] contains grayscale images with 28×28 pixel resolution. It contains 60000 train images and 10000 test images in total. To verify our effectiveness and save the processing time during this experiment, we randomly select 1000 samples as train data and 1000 samples as test data from the full train set and full test set(Fig. 4(a)). We select the hyper-parameters of our methods, by 5-fold cross-validation on train set with the searching space detailed in subsection A. Then using the best parameters obtained by parameter selection procedure 1) - 3) in Subsection A, all the classification results on test set are reported in Table I.

TABLE I. MNIST SUBSET CLASSIFICATION ACCURACY (%) OF DIFFERENT METHODS ON DIFFERENT DIMENSIONS

Feature-Dim	300	500	1000	1500	2000
NN	87.00	87.00	87.00	87.00	87.00
PCA	86.80	87.00	–	–	–
FDA	91.50	91.40	–	–	–
MFA	88.60	88.60	–	–	–
RBM	90.70	90.90	90.50	90.60	90.40
FDA-RBM	**92.40**	92.00	92.20	91.70	92.40
MFA-RBM	92.10	**92.40**	**92.30**	92.30	**92.70**
hkMFA-RBM	92.00	92.00	**92.30**	**92.50**	92.20

"$-$" means that the Feature-Dim of corresponding methods can not be reduced to the dimensionality in Table I. " $-$ " has the same meaning in subsequent experiments. From the results in Table I, we can find that: compared with the subspace models (PCA, FDA, MFA), the proposed hybrid models (FDA-RBM, MFA-RBM, hkMFA-RBM) can achieve much better classification accuracies in different subspaces. Moreover, due to its nonlinear property, the proposed hybrid models can also increase the dimensionality. In all the dimensionalities, the proposed models also outperform the traditional unsupervised RBM model. This identify that: integrating supervised subspace criteria into RBM model can consistently improve the classification performance.

C. Pendigits Experiment

The pendigits data set is an online pen-based digit set (10 classes) [26] and contains four different feature representations. There are 7494 samples for training and 3498 samples for testing. These feature representations are:

- (a) dyn (D = 16), eight successive pen points on two-dimensional coordinate system;
- (b) sta16 (D = 256), 16 x 16 image bitmap representation formed by connecting the points in dyn representation with line segments;
- (c) sta8 (D = 64), 8 x 8 subsampled bitmap representation;
- (d) sta4 (D = 16), 4 x 4 subsampled bitmap representation.

In this part, we use the (b) 16 x 16 image bitmap representation (Fig. 4(b)) to verify our algorithms.

Similar to MNIST experiment, we search the best parameters on the train set by parameter selection procedure in Subsection A . The test accuracies with these optimized parameters are listed in Table II.

TABLE II. PENDIGITS CLASSIFICATION ACCURACY (%) OF DIFFERENT METHODS ON DIFFERENT DIMENSIONS

Feature-Dim	50	100	200	300	500
NN	85.22	85.22	85.22	85.22	85.22
PCA	89.42	86.45	85.65	–	–
FDA	92.65	92.51	90.94	–	–
MFA	**94.60**	93.65	93.45	–	–
RBM	91.77	92.11	92.37	92.05	92.48
FDA-RBM	93.25	93.74	92.94	92.68	93.62
MFA-RBM	93.88	94.17	**94.77**	**94.71**	**94.63**
hkMFA-RBM	93.74	**94.25**	94.17	94.00	94.10

As shown in Table II, FDA-RBM, MFA-RBM, and hkMFA-RBM again outperform RBM model significantly. For extracting high-dimensional features, the proposed hybrid models outperform the traditional subspace models (PCA, FDA, MFA). However, in the low dimensional subspace, MFA performs better. The reason is that: the inter-class and intra-class graph of MFA are constructed on all the training samples, while for MFA-RBM, they are only construct on a mini-batch of samples for the purpose of mini-batch based gradient descending learning. Therefore, MFA-RBM may lose some global information of data, while MFA can make use of much more information about the global boundary relationships of training data.

D. Caltech101 Silhouettes Experiment

Caltech101 Silhouettes [27] [28] is a data set based on the Caltech101 image annotations. Each image in the Caltech101 data set includes a high-quality polygon outline of the primary object in the scene. The Caltech101 Silhouettes data set is created by centering and scaling each outline and rendering it on a quadrate pixel image-plane. The outline is rendered as a filled, black polygon on a white background (Fig. 4(c)).

There are two versions of this data set: outlines rendered as 28×28 images and outlines rendered as 16×16 images. And the total number of train set and test set are: 6364 and 2307, respectively. To evaluate our algorithm, we use the first 20 classes (Fig. 5) of the 28×28 version, which contains 2364 train data and 1281 test data (we call it Caltech20 Silhouettes).

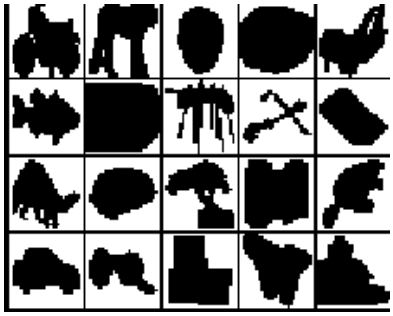

Fig. 5. The Caltech20 Silhouettes example images of 20 classes

TABLE III. CALTECH20 SILHOUETTES CLASSIFICATION ACCURACY (%) OF DIFFERENT METHODS ON DIFFERENT DIMENSIONS

Feature-Dim	300	500	1000	1500	2000
NN	77.13	77.13	77.13	77.13	77.13
PCA	77.67	77.75	–	–	–
FDA	80.48	79.16	–	–	–
MFA	79.86	78.77	–	–	–
RBM	80.33	80.56	80.25	80.17	80.41
FDA-RBM	81.11	80.80	80.80	81.11	80.80
MFA-RBM	**81.65**	81.03	81.03	81.11	80.56
hkMFA-RBM	81.19	**82.28**	**81.42**	**81.42**	**81.11**

We again search the best parameters by the procedure in Subsection A. However, one disappointing finding is that we can not get competitive results on this data set. In fact, caltech20 is sample unbalanced and has relative high dimension. Besides, searching space is compressed by optimizing only on important parameters with sparse gird in our parameter selection process. Thus, best performance is not guaranteed to be obtained. Instead of on train set, we search the optimized parameters on test set around the neighbors of the obtained parameters on train set, and use the results to compare with other methods. This kind of comparisons are also fair.

From Table III, we can get the same conclusion as previous experiments, which further verify the effectiveness of the proposed hybrid models.

VII. CONCLUSIONS AND FUTURE WORK

In this paper, the supervised subspace criteria (FDA, MFA, hkMFA) are proposed as regularization terms for the unsupervised RBM model to extract useful features for classification. The proposed hybrid algorithms impose discriminative constraint on the hidden layer of RBM during the contrastive divergence (CD) training process so that the model parameters can be updated both discriminatively and generatively. All the models are optimized based on mini-batch gradient descending. In the learning process, different criteria (FDA, MFA, hkMFA) are used to construct the within and between class adjacent matrices on mini-batch data, and then the gradients w.r.t. RBM and subspace criteria are calculated and combined for the

updating of the model parameters. This optimization process is as efficient as the traditional RBM model. Compared with the traditional subspace models, the proposed hybrid models are essentially nonlinear and can increase the dimensionality, while the traditional subspace models can only reduce dimensionality due to linear transformation. Experimental results verify that the proposed hybrid models can outperform both RBM model and the counterpart subspace models (FDA, MFA).

In the future, we will extend the input variable from binary to gaussian. On the other hand, the supervised subspace regularized RBM models will also be used as building-blocks to construct deep neural networks for further improving the accuracy.

REFERENCES

[1] G.E. Hinton and T.J. Sejnowski, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* Chapter 7 Learning and Relearning in Boltzmann Machines, pp. 282-317, MIT Press, USA, 1986.

[2] G.E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, pp. 1771-1800, 2002.

[3] G.E. Hinton and R.R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, pp. 504-507, 2006.

[4] R.R. Salakhutdinov and G.E. Hinton, "Deep Boltzmann machines," *Proceedings of The Twelfth International Conference on Articial Intelligence and Statistics*, vol. 5, pp. 448-455, 2009.

[5] Y. Bengio, P. Lamblin, D. Popovici and H. Larochelle, "Greedy layer-wise training of deep networks," *Advances in Neural Information Processing Systems*, vol. 19, pp. 153-160, 2007.

[6] R.R. Salakhutdinov and G.E. Hinton, "Semantic hashing," *International Journal of Approximate Reasoning*, vol. 50, pp. 969-978, 2009.

[7] H. Larochelle, Y. Bengio, J. Louradour and P. Lamblin, "Exploring strategies for training deep neural networks," *Journal of Machine Learning Research*, vol. 1, pp. 1-40, 2009.

[8] M. Ranzato, C. Poultney, S. Chopra and Y. LeCun, "Efficient learning of sparse representations with an energy-based model," *Advances in Neural Information Processing Systems*, pp. 1137-1144, 2006.

[9] H. Lee, C. Ekanadham and A. Y. Ng, "Sparse deep belief net model for visual area V2," *Advances in Neural Information Processing Systems*, pp. 873-880, 2008.

[10] V. Nair and G.E. Hinton, "3D object recognition with deep belief nets," *Advances in Neural Information Processing Systems*, pp. 1339-1347, 2009.

[11] H. Larochelle and Y. Bengio, "Classification using discriminative restricted Boltzmann machines," *Proceedings of the 25th international conference on Machine learning*, pp. 536-543, 2008.

[12] A. Stuhlsatz, J. Lippel and T. Zielke, "Feature extraction with deep neural networks by a generalized discriminant analysis," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, pp. 596-608, 2012.

[13] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, pp. 1373-1396, 2003.

[14] J.B. Tenenbaum, V. De Silva and J.C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319-2323, 2000.

[15] S.T. Roweis and L.K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323-2326, 2000.

[16] S. Yan, D. Xu, B. Zhang, H.J. Zhang, Q. Yang and S. Lin, "Graph embedding and extensions: a general framework for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 40-51, 2007.

[17] M.K. Wong and M. Sun, "Deep learning regularized Fisher mappings," *IEEE Transactions on Neural Networks*, vol. 22, pp. 1668-1675, 2011.

[18] J. Weston, F. Ratle, H. Mobahi and R. Collobert, *Neural Networks: Tricks of the Trade.* Chapter 26 Deep learning via semi-supervised embedding, pp. 639-655, Springer Berlin Heidelberg, 2012.

[19] R. Salakhutdinov and G.E. Hinton, "Learning a nonlinear embedding by preserving class neighbourhood structure," *International Conference on Artificial Intelligence and Statistics*, pp. 412-419, 2007.

[20] W. Yu, G. Zeng, P. Luo, F. Zhuang, Q. He and Z. Shi, "Embedding with Autoencoder Regularization," *Machine Learning and Knowledge Discovery in Databases*, pp. 208-223, 2013.

[21] M. Sugiyama, "Local Fisher discriminant analysis for supervised dimensionality reduction," *Proceedings of the 23rd international conference on Machine learning*, pp. 905-912, 2006.

[22] R.A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179-188, 1936.

[23] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* Boston Academic Press, Inc, 1990.

[24] P. Belhumeur, J. Hespanha and D. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, pp. 711-720, 1997.

[25] Y. LeCun and C. Cortes, "The MNIST database of handwritten digits," Available: http://yann.lecun.com/exdb/mnist.

[26] F. Alimoglu and E. Alpaydin, "Combining multiple representations and classifiers for pen-based handwritten digit recognition," *Proceedings of the Fourth International Conference on Document Analysis and Recognition*, vol. 2, pp. 637-640, 1997.

[27] B.M. Marlin, K. Swersky, B. Chen and N.D. Freitas, "Inductive principles for restricted Boltzmann machine learning," *Proc. International Conference on Artificial Intelligence and Statistics*, pp. 509-516, 2010.

[28] http://people.cs.umass.edu/ marlin/data.shtml.

[29] D. Cai, X.F. He and J.W. Han, "SRDA: an efficient algorithm for large scale discriminant analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, pp. 1-12, 2008.

[30] D. Cai, "Spectral Regression: A Regression Framework for Efficient Regularized Subspace Learning," PhD Thesis, UIUC, Computer Science, 2009.

[31] D. Cai, X.F. He, Y.X. Hu, J.W. Han and T. Huang, "Learning a Spatially Smooth Subspace for Face Recognition," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-7, 2007.

[32] G.E. Hinton, "A practical guide to training restricted Boltzmann machines," *Momentum*, 2010.

[33] Y. Bengio, J.F. Paiement, P. Vincent, O. Delalleau, N. Le Roux and M. Ouimet, "Out-of-sample extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral clustering," *Advances In Neural Information Processing Systems*, vol. 16, pp. 177-184, 2004.