Robust Feature Extraction via Information Theoretic Learning

Xiao-Tong Yuan XTYUAN@NLPR.IA.AC.CN
Bao-Gang Hu HUBG@NLPR.IA.AC.CN
NLPR/LIAMA, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China

Abstract

In this paper, we present a *robust* feature extraction framework based on information-theoretic learning. Its formulated objective aims at simultaneously maximizing the Renyi's quadratic information potential of features and the Renyi's cross information potential between features and class labels. This objective function reaps the advantages in robustness from both redescending M-estimator and manifold regularization, and can be efficiently optimized via half-quadratic optimization in an iterative manner. In addition, the popular algorithms LPP, SRDA and LapRLS for feature extraction are all justified to be the special cases within this framework. Extensive comparison experiments on several real-world data sets, with contaminated features or labels, well validate the encouraging gain in algorithmic robustness from this proposed framework.

1. Introduction

In this paper, we study the classical feature extraction problem, with the particular emphases on algorithmic robustness to data outliers and label noises. The training sample set is assumed to be represented as a matrix $X = [x_1, ..., x_N] \in \mathbb{R}^{m\times N}$, where N is the sample number and m is the original feature dimension. The class label indicator information of the training data is denoted by the matrix $C = [c_1, ..., c_N] \in \mathbb{R}^{N_c\times N}$, where N_c is the number of classes and the elements of the indicator vector c_i are set to be 1 or 0, according to whether x_i is drawn from the jth class. In practice, the feature dimension (m) is usually very high and thus it is necessary and beneficial to transform the data from the original high-dimensional space to a low-dimensional one for alleviating the curse of dimensionality (Fukunnaga, 1991). The purpose of linear feature extraction is to search for a projection matrix $W \in \mathbb{R}^{m'\times m}$ that transforms $x_i \in \mathbb{R}^m$ into a desired low-dimensional representation $y_i \in \mathbb{R}^{m'}$, where $m' \ll m$ and $y_i = Wx_i$.

Typically, the projection matrix W is learnt by optimizing a criterion describing certain desired or undesired statistical or geometric properties of the data set. Different criterions lead to different kinds of linear feature exaction algorithms. Among them, Principal Component Analysis (PCA) (Joliffe, 1986) and Linear Discriminant Analysis (LDA) (Fukunnaga, 1991) have been the two most popular ones owing to their simplicity and effectiveness. Another popular technique called Locality Preserving Projections (LPP) (He & Niyogi, 2004) has been proposed for linear feature extraction by preserving the local relationships within the data set. In (Yan et al., 2007), many classical linear feature extraction techniques are unified into a common framework known as Graph Embedding. To avoid the high time and memory usage associated with eigenvalue decomposition in LDA, the Spectral Regression Discriminant Analysis (SRDA) (Cai et al., 2008) was proposed based on ridge regression.

As these linear feature extraction methods are applied to realistic problems, where the amount of training data is large, it becomes impractical to manually verify whether all the data is "good". Taking image data as an example, the training data may contain undesirable artifacts due to image occlusion (e.g. a hand in front of a face), illumination (e.g. specular reflections), or image noise (e.g. from scanning archival data). We view these artifacts as statistical outliers (Huber, 1981). At the same time, for supervised learning, mislabeling of training data (e.g. confusing handwritten digit "3" with "8") may occur and deteriorate the performance of the learnt model. Therefore, the feature extraction techniques that can robustly derive low-dimensional

Appearing in *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009.

subspace from noisy data and labels is of particular interest in practice.

In this work, we present a novel feature extraction framework, called Renyi's Entropy Discriminative Analysis (REDA), towards algorithmic robustness to both data outliers and label noises via the formulation based on information-theoretic learning (ITL). The data set X is transformed into an N_c-dimensional feature space with the aim of maximizing an objective function related to the Renyi's entropy of the data features and the Renyi's cross-entropy between features and labels. The formulated problem can be viewed as a redescending M-estimator (Huber, 1981) of the SRDA with manifold regularization (Belkin et al., 2006), thus the REDA and robust statistics are well bridged. By utilizing the well known half-quadratic optimization technique (Rockfellar, 1970), the proposed objective function can be maximized in an iterative manner with theoretically provable convergence. In addition, for each iteration, the sub-problem is reduced into a LPP, SRDA or Laplacian Regularized Least Squares (LapRLS) (Belkin et al., 2006) problem, according to the values of the tunable trade-off parameter. The appealing characteristics of this proposed framework are summarized as follows: (1) Robust versions of LPP, SRDA and LapRLS can be derived within the proposed REDA framework, which helps users select a proper model according to given conditions; (2) Based on non-parametric Renyi's entropy estimation, REDA is not subject to any data distribution assumption; and (3) REDA can be efficiently solved via existing optimization techniques.

1.1. Related Works

The ITL based feature extraction has been extensively studied. In (Jenssen et al., 2006), a kernel transformation technique based on the idea of maximum entropy preservation was proposed for unsupervised feature extraction. The Informative Discriminative Analysis (Kaski & Peltonen, 2003) algorithm extracts a set of features by asymptotically maximizing mutual information that is computed based on a generative probabilistic model. In (Torkkola, 2003) and (Hild-II et al., 2006), feature extraction is conducted by directly maximizing the mutual information between the label and the features, with the entropy estimated by non-parametric Renyi's entropy.

The techniques for robust feature extraction have also attracted much attention recently. Algorithms like robust PCA (Torre & Black, 2001), robust LLE (Chang & Yeung, 2006) and robust Euclidean embedding (Cayton & Dasgupta, 2006) have been developed with sound theoretic justifications. As a complementarity to these works, our ITL motivated REDA framework implies the robust versions of the widely applied LPP, SRDA and LapRLS.

1.2. Paper Organization

The remainder of this paper is organized as follows. Section 2 introduces the non-parametric estimation of Renyi's quadratic/cross entropy. The problem formulation along with its robustness justification and optimization procedure are given in Section 3. Section 4 shows the experimental results and we conclude this work in Section 5.

2. Non-Parametric Renyi's Entropy

The Renyi's *quadratic* entropy of a probability density function $p(x)$ is defined as (Renyi, 1961)

$$H_2(x) = -\log\left(\int p^2(x)dx\right). \tag{1}$$

Suppose that the data set X is independently and identically drawn from $p(x)$, the following Gaussian kernel density estimation is then employed to estimate $p(x)$

$$\hat{p}(x) \propto \frac{1}{N}\sum_{i=1}^{N} g(x - x_i, \sigma)$$

where $g(x - x', \sigma) = \exp(-\|x - x'\|^2/\sigma^2)$. By substituting $p(x)$ with $\hat{p}(x)$ in (1) and after a series of simplifications, we arrive at the following non-parametric estimator for Renyi's quadratic entropy:

$$\begin{aligned}\hat{H}_2(X) &= -\log \hat{V}(X) + \text{const.} \\ \hat{V}(X) &= \sum_{i=1}^{N}\sum_{j=1}^{N} g(x_i - x_j, \sqrt{2}\sigma).\end{aligned}$$

Principe et al. (2000) named $\hat{V}(X)$ as the *information potential* (IP) of the set X, an analogy borrowed from physics for potential of group of interacting particles. Intuitively, the more regular set X is, the higher $\hat{V}(X)$ will be.

Following similar arguments, one can derive the equations for Renyi's *cross*-entropy between two sets X and X' as follows:

$$\begin{aligned}\hat{H}_2(X; X') &= -\log \hat{V}(X; X') + \text{const.} \\ \hat{V}(X; X') &= \sum_{i=1}^{N}\sum_{j=1}^{N} g(x_i - x'_j, \sqrt{2}\sigma).\end{aligned}$$

Intuitively, the cross IP $\hat{V}(X; X')$ reflects the extent of correlation between set X and X'.

Next, based on the above two IPs $\hat{V}(X)$ and $\hat{V}(X;X')$, we build the aforementioned robust linear feature extraction framework.

3. The Framework

3.1. Problem Formulation

We consider the projection matrix $W \in \mathbb{R}^{N_c \times m}$ that maps X into an $N_c \times N$ matrix $Y = WX$. The following criterion is used to encode the IP of feature Y and the cross IP between Y and the class label C,

$$E(W) = (1-\lambda)\hat{V}(WX) + \lambda\hat{V}(WX;C) \qquad (2)$$

where λ is a tunable trade-off parameter. The parameter W that maximizes $E(W)$ is desirable in the sense of minimizing the entropy of training set (reflected by the first unsupervised term), while separating training samples with different labels (reflected by the second supervised term). For a better statistical interpretation (see Section 3.2) of (2), we ignore the between class feature-label intersections contained in the term $\hat{V}(WX;C)$, thus the problem is finally formulated as:

$$\begin{aligned} W^* &= \arg\max_W \hat{E}(W) \\ &= \arg\max_W (1-\lambda)\sum_{i=1}^{N}\sum_{j=1}^{N} g(Wx_i - Wx_j, \sqrt{2}\sigma) \\ &\quad +\lambda\sum_{i=1}^{N} l_i g(Wx_i - c_i, \sqrt{2}\sigma) - \gamma\|W\|_2 \qquad (3) \end{aligned}$$

where l_i is the size of the class x_i belongs to, and term $\gamma\|W\|_2$ is the introduced Tikhonov regularization (with Frobenius norm) to avoid the possible overfitting to training data.

3.2. Robustness Justification

Let $\lambda = 1$ and $\gamma = 0$ in (3), we get

$$\begin{aligned} W^* &= \arg\max_W \sum_{i=1}^{N} l_i g(Wx_i - c_i, \sqrt{2}\sigma) \\ &= \arg\min_W \sum_{i=1}^{N} l_i \rho\left(\frac{Wx_i - c_i}{\sqrt{2}\sigma}\right) \qquad (4) \end{aligned}$$

where $\rho(u) = -\exp(-u^2)$. It is obvious that (4) is a robust *M-estimator* (Huber, 1981) formulation of the recently developed SRDA (Cai et al., 2008), with regressor X, observation C, regression parameter W and loss function $\rho(u)$. Moreover, $\rho(u)$ satisfies $\lim_{|u|\to\infty}\rho'(u) = 0$, thus it also belongs to the so called redescending M-estimators (Huber, 1981), which have in theory some special robustness properties, e.g., highest fixed design breakdown point (Mizera & Muller, 1999). Problem (4) is also known as a *correntropy* (Liu et al., 2007) optimization problem.

For general cases with $0 < \lambda < 1$, the second term in the objective function (3) remains a redescending M-estimator of SRDA. It can be seen from section 3.4.2 that the first term in (3) plays a role similar to manifold regularization used in LapRLS. Therefore, the proposed linear feature extraction formulation in (3) reaps the advantages of both robust statistics and manifold regularization.

3.3. Optimization

We apply the half-quadratic (HQ) optimization technique (Rockfellar, 1970) to solve problem (3).

3.3.1. Half Quadratic Optimization

Based on the *theory of convex conjugated functions* (Rockfellar, 1970), we can trivially derive the following proposition that forms the base to solve problem (3) in an HQ way.

Proposition 1 *There exists a convex function* $\varphi : \mathbb{R} \mapsto \mathbb{R}$*, such that*

$$g(x,\sigma) = \sup_{p\in\mathbb{R}^-}\left(p\frac{\|x\|^2}{\sigma^2} - \varphi(p)\right)$$

and for a fixed x*, the supremum is reached at* $p = -g(x,\sigma)$*.*

Now we introduce the following augmented objective function in an enlarged parameter space,

$$\begin{aligned} &\hat{F}(W,P,Q) \\ =\ &(1-\lambda)\sum_{i,j}\left(p_{ij}\frac{\|Wx_i - Wx_j\|^2}{2\sigma^2} - \varphi(p_{ij})\right) \\ &+\lambda\sum_i l_i\left(q_i\frac{\|Wx_i - c_i\|^2}{2\sigma^2} - \varphi(q_i)\right) \\ &-\gamma\|W\|_2 \end{aligned}$$

where the $N \times N$ matrix $P = [p_{ij}]$ and Q is diagonal with entity $Q(i,i) = q_i$ storing the auxiliary variables introduced in HQ analysis. According to the Proposition 1, we get immediately that for a fixed W, the following equation holds

$$\hat{E}(W) = \sup_{P,Q}\hat{F}(W,P,Q).$$

It follows that

$$\max_W \hat{E}(W) = \max_{W,P,Q}\hat{F}(W,P,Q), \qquad (5)$$

from which we can conclude that maximizing $\hat{E}(W)$ is equivalent to maximizing the augmented function $\hat{F}(W, P, Q)$ on the enlarged domain. Obviously, a local maximizer (W, P, Q) of $\hat{F}$ can be calculated in the following alternate maximization way:

$$p_{ij}^t = -g(W^{t-1}x_i - W^{t-1}x_j, \sqrt{2}\sigma), \quad (6)$$

$$q_i^t = -g(W^{t-1}x_i - c_i, \sqrt{2}\sigma), \quad (7)$$

$$W^t = \arg\max_W \mathrm{Tr}[WX(2(1-\lambda)L_p^t + \lambda LQ^t)X^T W^T - 2\lambda WXLQ^tC^T - \gamma WW^T], \quad (8)$$

where t means the t-th iteration, matrix L is diagonal with entity $L(i,i) = l_i$, Laplacian matrix $L_p^t = D_p^t - P^t$ where D_p^t is diagonal weight matrix whose entries are row sums of P^t, and $\mathrm{Tr}(\cdot)$ represents the matrix trace operation. We call this above three-step algorithm as *Renyi's Entropy Discriminant Analysis* (REDA) hereafter.

3.3.2. Convergence of REDA

Proposition 2 *Denote $\hat{F}^t = \hat{F}(W^t, P^t, Q^t)$, then the sequence $\{\hat{F}^t\}_{t=1,2,\ldots}$ generated by REDA algorithm converges.*

Proof We calculate

$$\hat{F}^t - \hat{F}^{t-1} = \left[\hat{F}(W^t, P^t, Q^t) - \hat{F}(W^{t-1}, P^t, Q^t)\right] + \left[\hat{F}(W^{t-1}, P^t, Q^t) - \hat{F}(W^{t-1}, P^{t-1}, Q^{t-1})\right].$$

According to Eq. (8) and the Proposition 1, both terms at the right side of above equal sign are non-negative. Therefore, the sequence $\{\hat{F}^t\}_{t=1,2,\ldots}$ is non-decreasing. It is easy to verify that both terms in $\hat{E}(W)$ are bounded above, and thus by Eq. (5) we get that $\hat{F}^t$ is also bounded. Consequently we can conclude that $\{\hat{F}^t\}_{t=1,2,\ldots}$ converges. □

3.4. Special Cases of REDA

We show that different setting of trade-off parameter λ will lead to special versions of REDA algorithm, which are highly related to the popular algorithms LPP, SRDA and LapRLS.

3.4.1. When $\lambda = 0$

Let $\lambda = 0$ and $\gamma = 0$, the calculation of Eq. (8) in REDA algorithm can be equivalently rewritten as

$$W^t = \arg\min_{WX(-D_p^1)X^TW^T = I} \mathrm{Tr}[WX(-L_p^t)X^TW^T]. \quad (9)$$

In this formulation, we introduce an extra constraint that $WX(-D_p^1)X^TW^T = I$, where I is an identity matrix, to remove arbitrary scaling and trivialness of solution, without breaking the convergence of algorithm. By initializing P^1 using the graph Laplacian (He & Niyogi, 2004), the calculation of W^1 is a standard LPP. When $t > 1$, (9) is a linear graph embedding problem with heat kernel similarity matrix $-P^t$ and constraint matrix $-D_p^1$, which can be efficiently solved via generalized eigenvalue decomposition method. We call this special version of our algorithm as REDA-LPP.

Basically, REDA-LPP is an unsupervised feature extraction algorithm. In practice, we may extend it into a supervised version by setting $p_{ij}^t = 0$, if $c_i \neq c_j$. Interestingly, the supervised REDA-LPP also implies the robustness against outliers. It is known that at each iteration t, graph embedding problem (9) aims to preserve on the set W^tX the sample pairwise similarity measured among the previous set $W^{t-1}X$. Typically, an outlier $W^{t-1}x_k$ is far away from the data cluster of its class and thus always receives low p_{kj}^t to $W^{t-1}x_j$ of the same class. Therefore, the outliers will have weaker influence on the estimation of W^t as t increases.

3.4.2. When $0 < \lambda \leq 1$

In this case, the Eq. (8) in REDA is calculated as

$$W^t = \lambda\left(X(2(1-\lambda)L_p^t + \lambda LQ^t)X^T - \gamma I\right)^{-1} XLQ^tC^T. \quad (10)$$

- When $\lambda = 1$, by initializing $q_i^1 = -1$, the calculation of W^1 is equivalent to SRDA. When $t > 1$, the auxiliary variable $-q_i^t$ gives the weight of (x_i, c_i) for the estimation of W^t via SRDA. We refer to this version of our algorithm as REDA-SRDA, which is the solution of M-estimator (4).

- When $0 < \lambda < 1$, it is easy to see that, at each iteration t, Eq. (10) is the solution of a LapRLS problem with graph similarity matrix $-P^t$ based on previous representation. Such an iterative LapRLS feature extraction method reaps both the the robustness of M-estimator and the advantage of manifold regularization. We call this version of our algorithm as REDA-LapRLS.

The connections of our REDA algorithm with those existing algorithms are summarized in Table 1.

3.5. Learning of Response C

In this work, we conventionally choose each column of the response C in (3) as class label indicator. Actually, as pointed out in (Cai et al., 2008) that C can be more generally learnt via some graph embedding

Table 1. Connections of REDA with existing algorithms.

Setting	Connections
$\lambda = 0$, $\gamma = 0$, $t = 1$	Standard LPP
$\lambda = 0$, $\gamma = 0$, $t > 1$	Robust extension for LPP
$\lambda = 1$, $t = 1$	Standard SRDA
$\lambda = 1$, $t > 1$	Robust extension for SRDA
$\lambda \in (0, 1)$, $t = 1$	Standard LapRLS
$\lambda \in (0, 1)$, $t > 1$	Robust extension for LapRLS

algorithms, e.g., LDA and LPP, with different dimensions m'. Specially, when $m' = N_c$, the learnt spectral response C by LDA is equivalent to the one used here.

3.6. Kernel Extension

Commonly, algorithm for linear feature extraction is computationally efficient for both projection matrix learning and final classification. However, its performance may degrade in cases with nonlinearly distributed data. A technique to extend methods for linear projections to nonlinear cases is to directly take advantage of the kernel trick. The intuition of the kernel trick is to map the data from the original input space to another higher dimensional Hilbert space as $\phi : X \mapsto Z$, and then perform the linear algorithm in this new feature space. This approach is well-suited to algorithms that only need to compute the inner product of data pairs $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$. Assuming that the projection matrix $W = A\Phi$, where $\Phi = [\phi(x_1), ..., \phi(x_N)]^T$ and $\mathcal{K}$ is the kernel Gram matrix with entity $\mathcal{K}(i, j) = k(x_i, x_j)$, we have the following kernelization of problem (3),

$$\begin{aligned} A^* &= \arg\max_A (1-\lambda) \sum_{i,j} g(A\mathcal{K}_i - A\mathcal{K}_j, \sqrt{2}\sigma) \\ &\quad + \lambda \sum_i l_i g(A\mathcal{K}_i - c_i, \sqrt{2}\sigma) - \gamma \|A\|_2, \end{aligned}$$

where $\mathcal{K}_i$ indicates the ith column vector of the kernel Gram matrix $\mathcal{K}$. Accordingly, we can derive the so called KREDA-LPP, KREDA-SRDA and KREDA-LapRLS algorithms for robust kernel-based feature extraction.

4. Experiments

To evaluate the robustness of different special versions of our proposed REDA algorithm, we systematically compare them with their traditional counterparts on several real-world data sets, with contaminated features or labels.

4.1. Data Sets

We use the Extended Yale Face Database B[1], the MNIST handwritten digit database[2] and the TDT2 document database[3] for performance evaluation. Here are some basic information about these three data sets.

Extended Yale Face Database B (YaleB) The YaleB database contains 16128 images of 38 human subjects under 9 poses and 64 illumination conditions. We use 64 near frontal face images for each individual in our experiment. The size of each cropped gray scale image is 32×32 pixels. For each individual, $N = (20, 30, 40)$ images are randomly selected for training (with $m = 1024$ and $N_c = 38$), and the rest are used for testing.

MNIST Handwritten Digits Database The MNIST database of handwritten digits has a training set A of 60,000 examples, and a test set B of 10,000 examples. The digits have been size-normalized and centered in a fixed-size (28×28) bilevel image. In our experiment, we use the digits $\{3, 8, 9\}$ which represent difficult visual discrimination problem. We take the $\{3, 8, 9\}$ digits in the first 10000 samples from set A as our training set and those in the first 10000 from set B as our test set. A random subset with $N = (100, 200, 300)$ samples per digit from the training set is selected for training (with $m = 784$ and $N_c = 3$).

TDT2 Document Database The TDT2 corpus consists of 11,201 on-topic documents which are classified into 96 semantic categories. We use the top 9 categories for our experimental evaluation. Each document is represented as a normalized term-frequency vector, with top 2000 words selected according to mutual information. For each category, $N = (30, 60, 100)$ documents are randomly selected for training (with $m = 2000$ and $N_c = 9$), and the rest are used for testing.

4.2. Experiment Design

We compare the following algorithms on the YaleB and MNIST data sets:

1. LPP and our REDA-LPP.

2. SRDA and our REDA-SRDA.

3. LapRLS and our REDA-LapRLS.

[1] `http://vision.ucsd.edu/~leekc/ExtYaleDatabase/ExtYaleB.html`

[2] `http://yann.lecun.com/exdb/mnist/`

[3] `http://www.nist.gov/speech/tests/tdt/tdt98/index.htm`

Table 2. Performance comparison on YaleB set. $\sigma = 0.47$, $\gamma = N$.

Methods	Classification Errors (mean ± std-dev %)								
	$N \times N_c = 20 \times 38$			$N \times N_c = 30 \times 38$			$N \times N_c = 40 \times 38$		
	$\eta = 0\%$	$\eta = 25\%$	$\eta = 50\%$	$\eta = 0\%$	$\eta = 25\%$	$\eta = 50\%$	$\eta = 0\%$	$\eta = 25\%$	$\eta = 50\%$
REDA-LPP	5.7	**8.9**±0.7	**13.1**±1.2	3.0	**4.8**±0.4	6.9±0.8	2.4	3.6±0.4	5.6±0.5
LPP	6.1	15.4±1.0	21.9±2.1	2.7	10.0±0.6	14.2±0.8	1.8	6.5±0.6	10.7±1.0
REDA-SRDA	4.7	9.1±0.7	13.3±0.9	1.8	5.6±0.6	**6.8**±1.1	1.8	**2.7**±0.2	6.2±1.1
SRDA	4.7	13.1±1.4	19.1±1.6	**1.7**	8.5±0.6	11.8±0.7	1.7	5.8±0.3	9.5±0.7
REDA-LapRLS	5.0	9.2±0.7	13.3±1.0	1.9	5.0±0.1	**6.8**±1.1	**1.1**	2.9±0.2	**5.5**±0.8
LapRLS	4.8	12.8±1.2	19.0±1.7	1.8	8.5±0.6	11.7±0.7	**1.1**	5.7±0.2	9.3±0.7
RLDA	**4.4**	12.6±0.1	18.2±2.0	**1.7**	8.0±0.5	11.6±0.4	1.2	5.3±0.3	9.2±0.5
Robust PCA	31.5	35.3±0.8	39.7±1.2	24.4	27.9±0.8	30.4±0.8	20.3	22.3±1.1	26.0±1.0

4. Regularized LDA (RLDA) (Friedman, 1989) as non-robust baseline.

5. Robust PCA (Torre & Black, 2001) as robust baseline.

On the TDT2 corpus, we compare the kernel extensions of the above algorithms. The second order polynomial kernel is used to construct Gram matrix $\mathcal{K}$.

As aforementioned, we aim to test the performance of the compared algorithms when training sets are contaminated by outliers or mislabeling, which are generated in the following artificial ways:

- For the YaleB data set, from each individual, we randomly select $\eta = (25\%, 50\%)$ training sample images and partially occlude in them some key facial features. See Figure 1 for some selected sample images with outliers.

- For MNIST and TDT2 data sets, from each training class, we randomly select $\eta = (25\%, 50\%)$ samples and then label each of them as one of the other classes with equal probabilities.

Figure 1. Selected sample images without and with artificial outliers in the YaleB set. Top row: clean images; Middle row: outliers by forehead and eyes occlusion; Bottom row: outliers by nose and mouth occlusion.

To evaluate the discriminability of the learnt subspace, the classification error from the *nearest center* classifier on test set is finally used as the evaluation metric.

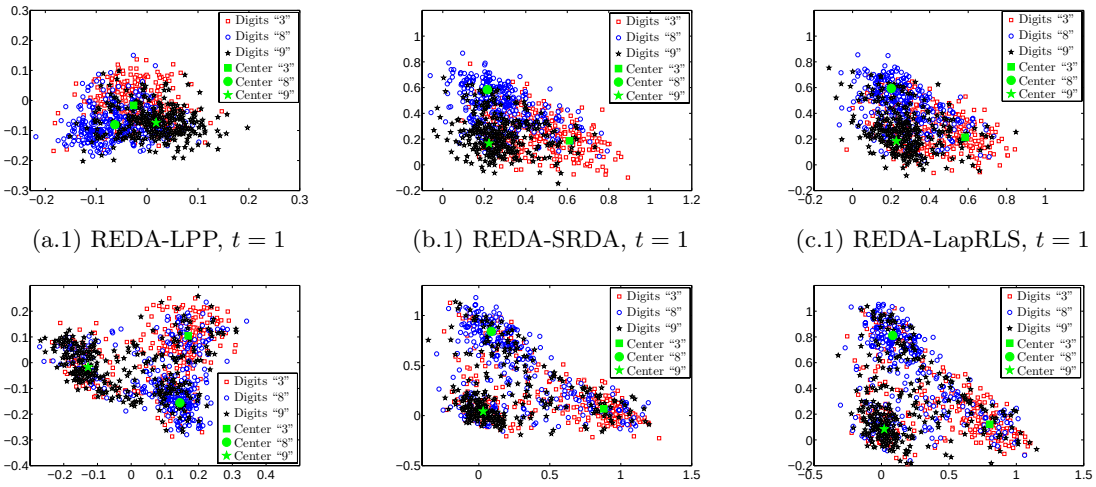
4.3. Results

4.3.1. Illustration of Robustness

To visualize the robustness of the proposed REDA-LPP, REDA-SRDA and REDA-LapRLS, we apply them on the MNIST set. In this example, each digit class is of size $N = 300$ with $\eta = 50\%$ training samples being randomly mislabeled as the other digits. We set $\lambda = 0.99$ in REDA-LapRLS throughout the experiments. When $t = 1$, the standard LPP (Figure 2(a.1)), SRDA (Figure 2(b.1)) and LapRLS (Figure 2(c.1)) all perform poorly to discriminate classes in the learnt subspace due to mislabeling. When convergence is attained at $t = 6$ for all these three REDA algorithms, much more discriminative results are achieved, as can be seen in Figure 2(a.2), 2(b.2)&2(c.2). The enhanced discriminability of our algorithms on dirty data also leads to the significant improvement of classification performance, as can be seen from the quantitative results provided in next sub-section.

4.3.2. Quantitative Results

Tables 2∼4 list the test errors of compared algorithms on the three data sets separately. For each given training size N and outlier (mislabeling) percentage $\eta > 0$, the test error mean and standard deviation are estimated according to 50 times of running under random outlier (mislabeling) generation. When training set is clean, it can be seen that the test performance is comparable among our REDA methods and their related traditional methods. This is because without apparent outliers, only one single regression cluster appears in the data, thus the robust statistics does not help to improve the performance of parameter estimation and classification. When outliers or mislabeling are introduced in training sets, the robustness of our REDA methods functions and much lower test errors are consistently achieved by our methods compared to their non-robust counterparts, as well as the RLDA and robust PCA. Interestingly, we observe that for the mislabeling cases in MNIST and TDT2 data sets, when training set size N is relatively large (see the right



(a.1) REDA-LPP, $t = 1$ (b.1) REDA-SRDA, $t = 1$ (c.1) REDA-LapRLS, $t = 1$

(a.2) REDA-LPP, $t = 6$ (b.2) REDA-SRDA, $t = 6$ (c.2) REDA-LapRLS, $t = 6$

Figure 2. Feature extraction results of a MNIST training set by REDA-LPP, REDA-SRDA and REDA-LapRLS. Here, the first two dimensions of output features are plotted for visualization. Each class center is robustly estimated via iteratively re-weighted least squares (IRLS).This figure is better viewed in color and please see text for the detailed descriptions.

three columns of Table 3&4), the test errors by our REDA methods are relatively stable as the η increases from 0% to 50%. We also observe that the unsupervised robust PCA is insensitive to label noise as shown on these two data sets. In all our experiments, the convergence of REDA can be attained after less than 10 iterations.

4.3.3. Parameter Selection for REDA

We estimate the kernel scale parameter σ by adopting the technique of simultaneous regression-scale estimation (Mizera & Muller, 2002). γ is another essential parameter in REDA-SRDA and REDA-LapRLS algorithms that controls the smoothness of M-estimator. The reported results in this paper are obtained under $\gamma = N$, while our numerical observation shows that REDA performs well over a large range of γ.

5. Conclusions and Future Work

In this paper, a robust feature extraction framework was derived by maximizing an objective function motivated by Renyi's quadratic and cross entropy. As analyzed, the main advantage of this proposed framework lies in its robustness against training outliers for both features and labels. We proposed to utilize the half-quadratic optimization technique to solve the formulated optimization problem in an iterative manner. At each iteration the problem was reduced to a quadratic optimization problem which can be efficiently optimized. The connections between our proposed framework and several existing popular feature extraction algorithms were highlighted. One interesting future research direction is to study REDA further within the settings of robust semi-supervised learning and robust transfer learning.

Acknowledgement

The authors would like to thank Dr. Shuicheng Yan for reading an earlier version of this manuscript and his valuable feedbacks for refinement. This work was supported in part by NSF of China (No. 60275025) and MOST of China (No. 2007DFC10740).

References

Belkin, M., Niyogi, P., & Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 2399–2434.

Cai, D., He, X., & Han, J. (2008). Srda: An efficient algorithm for large-scale discriminant analysis. *IEEE Transactions on Knowledge and Data Engineering, 20(1)*, 1–12.

Cayton, L., & Dasgupta, S. (2006). Robust euclidean embedding. *International Conference on Machine Learning* (pp. 169–176).

Chang, H., & Yeung, D.-Y. (2006). Robust locally linear embedding. *Pattern Recognition, 39(6)*, 1053–1065.

Table 3. Performance comparison on MNIST set. $\sigma = 0.83$, $\gamma = N$.

Methods	Classification Errors (mean ± std-dev %)								
	$N \times N_c = 100 \times 3$			$N \times N_c = 200 \times 3$			$N \times N_c = 300 \times 3$		
	$\eta = 0\%$	$\eta = 25\%$	$\eta = 50\%$	$\eta = 0\%$	$\eta = 25\%$	$\eta = 50\%$	$\eta = 0\%$	$\eta = 25\%$	$\eta = 50\%$
REDA-LPP	7.4	9.0±0.7	14.4±0.8	6.4	**7.1**±0.4	**8.2**±0.6	5.4	7.1±0.4	7.7±0.2
LPP	7.6	11.5±1.1	20.5±3.3	6.5	9.0±0.8	13.1±1.9	5.4	8.3±0.2	12.3±2.0
REDA-SRDA	7.9	**8.6**±0.5	**12.0**±1.7	6.2	7.2±0.4	9.2±1.2	5.5	**6.4**±0.2	7.6±0.6
SRDA	7.8	11.8±1.2	23.0±3.4	6.1	8.6±0.9	15.7±1.6	5.3	7.5±0.1	15.0±2.1
REDA-LapRLS	8.0	9.1±0.6	14.9±2.1	6.3	7.3±0.5	9.9±1.3	5.5	6.5±0.2	**7.3**±0.3
LapRLS	7.8	11.9±1.3	23.4±3.5	**5.9**	8.6±0.9	16.1±1.7	5.3	7.6±0.2	15.3±2.1
RLDA	**7.3**	11.4±1.1	23.0±3.4	6.0	8.3±0.7	15.8±1.6	**5.2**	7.4±0.3	14.9±2.0
Robust PCA	12.4	13.0±1.8	15.9±3.7	11.2	12.4±0.6	13.0±1.1	10.9	11.4±0.4	11.9±1.4

Table 4. Performance comparison on TDT2 corpus. $\sigma = 0.64$, $\gamma = N$.

Methods	Classification Errors (mean ± std-dev %)								
	$N \times N_c = 30 \times 9$			$N \times N_c = 60 \times 9$			$N \times N_c = 100 \times 9$		
	$\eta = 0\%$	$\eta = 25\%$	$\eta = 50\%$	$\eta = 0\%$	$\eta = 25\%$	$\eta = 50\%$	$\eta = 0\%$	$\eta = 25\%$	$\eta = 50\%$
KREDA-LPP	9.5	**10.5**±0.5	**12.1**±0.4	8.1	8.2±0.3	**8.6**±0.8	7.0	6.9±0.3	**6.8**±0.7
KLPP	9.3	12.3±1.2	19.2±2.5	8.5	10.8±1.2	14.0±0.8	6.9	9.1±0.3	12.9±1.3
KREDA-SRDA	9.2	**10.5**±0.4	12.8±1.1	**7.6**	**8.0**±0.5	9.0±0.7	6.4	**6.7**±0.1	6.9±0.6
KSRDA	9.4	13.3±1.4	18.9±2.2	8.0	10.8±1.3	14.1±1.9	6.6	8.0±0.5	10.8±1.2
KREDA-LapRLS	**9.1**	**10.5**±0.5	12.8±1.0	**7.6**	**8.0**±0.5	9.0±0.7	**6.3**	6.8±0.1	7.1±0.7
KLapRLS	9.4	12.7±1.1	19.0±2.2	8.0	10.8±1.3	14.2±0.9	**6.3**	8.0±0.4	10.8±1.2
KRLDA	9.2	12.4±0.8	19.5±2.4	8.3	10.6±0.9	14.1±0.9	6.6	9.4±0.4	13.2±1.4
Robust KPCA	13.2	14.8±1.3	15.6±1.7	10.3	10.1±0.6	10.7±1.2	8.4	8.7±0.3	8.5±0.4

Friedman, J. (1989). Regularized discriminative analysis. *Journal of American Statistical Association, 84(405)*, 165–175.

Fukunnaga, K. (1991). *Introduction to statistical pattern recognition.* Academic Press.

He, X., & Niyogi, P. (2004). Locality preserving projections. *Advances in Neural Information Processing Systems 16.* Cambridge, MA: MIT Press.

Hild-II, K., Erdogmus, D., Torkkola, K., & Principe, C. (2006). Feature extraction using information-theoretic learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(9)*, 1385–1392.

Huber, P. (1981). *Robust statistics.* Wiley.

Jenssen, R., Eltoft, T., Girolami, M., & Erdogmus, D. (2006). Kernel maximum entropy data transformation and an enhanced spectral clustering algorithm. *Advances in Neural Information Processing Systems 19* (pp. 633–640). Cambridge, MA: MIT Press.

Joliffe, I. (1986). *Principal component analysis.* Springer-Verlag.

Kaski, S., & Peltonen, J. (2003). Informative discriminant analysis. *International Conference on Machine Learning* (pp. 329–336).

Liu, W., Pokharel, P. P., & Principe, J. C. (2007). Correntropy: Properties and applications in non-gaussian signal processing. *IEEE Transactions on Signal Processing, 55(11)*, 5286–5298.

Mizera, I., & Muller, C. (1999). Breakdown points and variation exponents of robust m-estimators in linear models. *Annals of Statistics, 27*, 1164–1177.

Mizera, I., & Muller, C. (2002). Breakdown points of cauchy regression-scale estimators. *Statistics and Probability Letters, 57*, 79–89.

Principe, J., Xu, D., & Fisher, J. (2000). Information theoretic learning. *Unsupervised Adaptive Filtering.* New York: Wiley.

Renyi, A. (1961). On measures of information and entropy. *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability* (pp. 547–561).

Rockfellar, R. (1970). *Convex analysis.* Princeton Press.

Torkkola, K. (2003). Feature extraction by nonparametric mutual information maximization. *Journal of Machine Learning Research, 3*, 1415–1438.

Torre, F., & Black, M. (2001). Robust principal component analysis for computer vision. *International Conference on Computer Vision* (pp. 362–369).

Yan, S., Xu, D., Zhang, B., Zhang, H., Yang, Q., & Lin, S. (2007). Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(1)*, 40–51.