

# Relevant Linear Feature Extraction Using Side-information and Unlabeled Data

Fei Wu

Department of Automation  
Tsinghua University, Beijing, China  
wufei98@mails.tsinghua.edu.cn

Yonglei Zhou

Department of Automation  
Tsinghua University, Beijing, China  
zhouyonglei98@mails.tsinghua.edu.cn

Changshui Zhang

Department of Automation  
Tsinghua University, Beijing, China  
zcs@mail.tsinghua.edu.cn

## Abstract

*“Learning with side-information” is attracting more and more attention in machine learning problems. In this paper, we propose a general iterative framework for relevant linear feature extraction. It efficiently utilizes both the side-information and unlabeled data to enhance gradually algorithms’ performance and robustness. Both good relevant feature extraction and reasonable similarity matrix estimation can be realized. Specifically, we adopt Relevant Component Analysis (RCA) under this framework and get the derived Iterative Self-Enhanced Relevant Component Analysis (ISERCA) algorithm. The experimental results on several data sets show that ISERCA outperforms RCA.*

## 1 Introduction

The performance of many learning and data mining algorithms, such as K-means clustering and nearest-neighbor searching, depends critically on the feature extraction criterion over the input space.

Recently, a new focus called “learning with side-information” is proposed in some literatures [1][4][7][8]. Side-information represents some equivalence constraint between pair of samples, indicating whether the two samples originate from the same but *unknown* category (positive constraint) or from two different categories (negative constraint). We use “labeled data” to denote the samples involved in the given side-information.

In this paper, we concern the problem of relevant linear feature extraction with side-information. Some related works have appeared in [1][8]. In [8], Xing et. al. propose to learn a semi-positive matrix using side-information. The objective is to minimize the sum of distances in the associ-

ated feature space between each positively constrained pair. The Relevant Component Analysis (RCA) algorithm, presented by Shental et. al. in [1], uses the positive constraints to reduce irrelevant variabilities of data while amplifying relevant variabilities. Results in [1] and [8] show that in many cases both these algorithms have better performance than some classical methods, such as PCA. Nevertheless, they still can be enhanced at some aspects. For example, they utilize only the given side-information (specifically, RCA only uses the positive constraints), ignoring the underlying information contained in large amounts of unlabeled data. This weakens the algorithms’ performance and robustness, especially when the given side-information is insufficient. In many cases we do have little knowledge about data - data is too much to be labeled, or the label information is too expensive to obtain.

In this paper, we propose a general iterative framework for relevant feature extraction. With effective utilization of both the labeled and unlabeled data in an iterative way, it can enhance algorithms’ performance and robustness. Specifically, we introduce RCA into this framework and get a derived algorithm named Iterative Self-Enhanced Relevant Component Analysis (ISERCA). Experimental results on several data sets show that our ISERCA obtains an obvious performance and robustness improvement compared with RCA, especially when the given side-information is insufficient.

The rest of this paper is organized as follows. In section 2, we propose a general iterative framework on how to extract relevant linear features with side-information and unlabeled data. In section 3, we will formulate the derived ISERCA algorithm. Some experiments and results are given in section 4, followed by conclusions in section 5.

## 2 An iterative framework for relevant linear feature extraction

Suppose we have some data set  $\chi = \{x_i\}_{i=1}^N \subseteq R^D$  sampled independently from  $K$  discrete sources. Let  $\chi$  and  $\chi \times \chi$  denote the original data space and the product space respectively. Thus we can define a similarity function on the product space,  $s(\cdot, \cdot) : \chi \times \chi \rightarrow [0, 1]$ , which represents the similarity score of each pair of data points.

Define the similarity matrix as  $S = \{s_{ij} = s(x_i, x_j)\}_{i,j=1}^N$ . The idea of the iterative framework is intuitive. If more elements of  $S$  are accurately estimated, more side-information could be obtained, then the extracted features could be more relevant. Meanwhile, features which are more relevant can lead to a more accurate estimation of  $S$ . We can perform the feature extraction with the current similarity matrix and similarity matrix estimation in the new feature space by turns, to gradually achieve better results.

Gaussian Mixture Model (GMM) can be used to estimate  $S$ . Its parameter set is evaluated by the constrained-EM algorithm presented in [4], for the side-information can be incorporated into this procedure. In the process of feature extraction, both RCA and Xing's method can be adopted. Here only RCA is considered for linear feature extraction. The derived algorithm named Iterative Self-Enhanced Relevant Component Analysis (ISERCA) will be detailed in the next section.

## 3 ISERCA

Based on the iterative framework above, the idea of ISERCA is very straightforward: RCA is used to select a linear feature space using the current positive constraints. Then a generative model is learned in the current feature space to augment the positive constraints. Iterate the process until some target is hit.

### 3.1 Linear feature extraction by RCA

Suppose we have some data set  $\chi = \{x_i\}_{i=0}^N \subseteq R^D$ , and some side-information. The side-information can be divided into two subsets: positive constraints set  $PC = \{(x_i, x_j) | s_{i,j} = 1\}$  and negative constraints set  $NC = \{(x_i, x_j) | s_{i,j} = 0\}$ . Let a chunklet  $c_i$  denote a small subset of data points that are known to belong to a single but unknown source. The chunklets can be obtained by applying the transitive closure to  $PC$ .

Briefly, the RCA algorithm can be described as follows:

1. Based on  $PC$  get the chunklet set  $C = \{c_i\}_{i=1}^{|C|}$ , which satisfy  $\chi = \bigcup_{i=1}^{|C|} c_i$ . The data points in the  $j$ th chunklet are represented as  $\{x_i^j\}_{i=1}^{|c_j|}$ .

2. Compute the weighted within-chunklet scatter matrix by

$$\hat{C} = \frac{1}{N} \sum_{j=1}^{|C|} \sum_{i=1}^{|c_j|} (x_i^j - m^j)(x_i^j - m^j)^T \quad (1)$$

where  $m^j$  is the mean of the  $j$ th chunklet.

3. Compute the whitening transformation matrix  $W = \hat{C}^{-1/2}$ . Convert data point  $x$  into the feature space as follows:

$$y = Wx \quad (2)$$

Further work can be continued in the feature space based on Euclidean metric.

RCA has shown good performance in many applications. Nevertheless, it still can be enhanced. As described in the algorithm, the objective of RCA is to use the within-chunklet scatter matrices to approximate the covariance matrices and to make the covariance matrices of all classes spherical. When the positive constraint set is relatively small, commonly in practice, the within-chunklet scatter matrices are unlikely to approximate the true covariance matrices well. Inevitably the performance and robustness of RCA will decrease.

### 3.2 Similarity matrix estimation

We use Gaussian Mixture Model to approximate the distribution of data:

$$p(x|\Theta) = \sum_{l=1}^K \alpha_l N(x; \mu_l, \Sigma_l) \quad (3)$$

where,  $\alpha_l$  denotes the weight of each component, and  $N(x; \mu_l, \Sigma_l)$  represents a normal distribution with the mean  $\mu_l$  and the covariance  $\Sigma_l$ .

Define the probability for  $x_i$  sampled from the  $j$ th Gaussian component as

$$O_i^j = \frac{\alpha_j N(x_i; \mu_j, \Sigma_j)}{\sum_{l=1}^K \alpha_l N(x_i; \mu_l, \Sigma_l)} \quad (4)$$

Thus a hypothesis over the product space  $h : \chi \times \chi \rightarrow [0, 1]$  can be constructed straightforward:

$$h(i, j) = \sum_{l=1}^K O_i^l \cdot O_j^l \quad (5)$$

It corresponds to the probability that the two data points  $x_i$  and  $x_j$  originate from the same class. It can be regarded as the estimation of  $s_{i,j}$ .

Constrained-EM is used to estimate the parameter set of the GMM as proposed in [4]. It differs from the standard

EM only in the ‘E’ step - the sum is taken only over assignments which comply with the given side-information instead of over all possible assignments.

The obtained hypothesis  $h(\cdot, \cdot)$  as in equation (5) is always not reliable enough. It can be seen as a weak learner over the product space. As proposed in [3], boosting is an efficient method which linearly combines weak learners into a strong learner. Hence we use a similar boosting scheme as the one in [3] to construct a combined strong learner as follows:

$$L_M(i, j) = \sum_{l=1}^M \alpha_l h_l(i, j) \quad (6)$$

where,  $h_l(i, j)$  is the  $l$ th weaker learner and  $\alpha_l$  denotes its weight.

More details about the boosting procedure can be found in [3] [5]. In this procedure information implied by both labeled and unlabeled data is sufficiently considered.

Finally the output of the boosting procedure is normalized as follows:

$$\tilde{L}_M(i, j) = \frac{L_M(i, j)}{\sum_{l=1}^M \alpha_l} \in [0, 1]. \quad (7)$$

We regard  $\tilde{L}_M(i, j)$  as an estimation of the similarity  $s_{i,j}$ .

### 3.3 ISERCA algorithm

Now we can formulate the ISERCA algorithm.

1. Given data set  $\chi$  and side-information, build chunklets  $C$ . Define two sub data sets:

$$\hat{P} = \{x_i | x_i \text{ is evolved in positive constraints}\}$$

$$\hat{N} = \{x_i | x_i \text{ is evolved in negative constraints}\}$$

Then compute two reference variables: positive constraint ratio  $pr = \frac{|\hat{P}|}{N}$ , and negative constraint ratio  $nr = \frac{|\hat{N}|}{N}$ .

2. If  $pr$  is bigger than the preset threshold  $\alpha_{pr}$  (usually set as 0.1), transform  $\chi$  into a feature space  $\psi_f$  by normal RCA. Otherwise denote the original data space as  $\psi_f$ .

3. In the product space  $\psi_f \times \psi_f$ , the boosting procedure described above is carried out to build a strong learner. Its output  $\tilde{L}_M(i, j)$  is taken as the estimation of the similarity  $s_{i,j}$ .

4. Add pair data points with bigger  $s_{i,j}$  to the original positive constraint set and update the chunklets. This operation must comply with the given side-information, specifically the negative constraints.

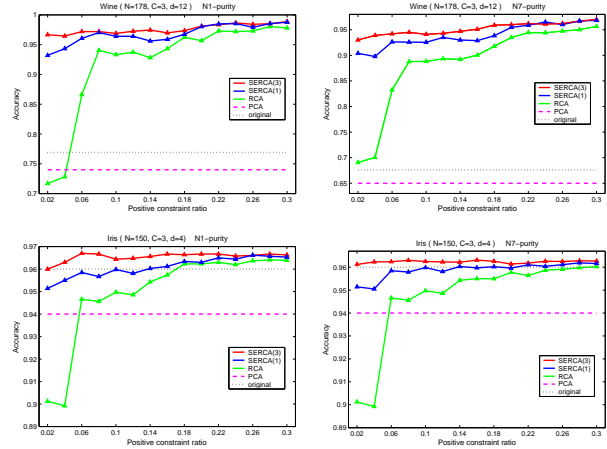
5. Using updated chunklets, perform RCA in  $\psi_f$  to get a new feature space  $\tilde{\psi}_f$ . If the target is hit, terminate and

output  $\tilde{\psi}_f$ . Otherwise denote  $\tilde{\psi}_f$  as  $\psi_f$  and turn to step 3.

The proposed ISERCA utilizes not only the given positive constraints but also the information implied by negative constraints and unlabeled data to extract features in an iterative scheme. The following experimental results show that ISERCA is more efficient and robust than RCA.

## 4 Experiments and results

To evaluate the performance of our ISERCA algorithm, we tested it on the *wine* and *iris* data sets from UCI repository. Mean neighbor purity  $p(n)$  is the evaluation criterion. Specifically, for each data point  $\tilde{x}_i$  we find its  $n$  nearest neighbors based on Euclidean metric in the feature space. If  $m$  of these  $n$  neighbors originate from the same category as  $\tilde{x}_i$  does, its  $n$  neighbor purity is evaluated as  $p_i(n) = m/n$ .  $p(n)$  is the mean of all  $p_i(n)$  (for  $i = 1, \dots, N$ ). If the difference of the similarity matrices is small enough among several consecutive iterations, the algorithm is terminated. In most of our experiments 3 iterations were enough. All the experiments were performed 20 times for each run with the same positive constraint ratio  $pr$  and negative constraint ratio  $nr$ . We calculated the mean  $\bar{p}(n)$  and the covariance of 20 runs’ results and compared them in the following spaces: original input space, feature space by PCA, feature space by RCA, feature space by ISERCA with one iteration, and feature space by ISERCA with three iterations.



**Figure 1.**  $\bar{p}(1)$  and  $\bar{p}(7)$  computed in five spaces over wine and iris data set. The horizontal axis denotes the positive constraint ratio, and the vertical axis denotes the mean neighbor purity. Negative constraint ratio is fixed as 0.1.

Specifically, we computed  $\bar{p}(1)$  and  $\bar{p}(7)$  respectively to

$pr$	0.020	0.080	0.140	0.200	0.260
RCA	0.058	0.034	0.033	0.011	0.012
SERCA(1)	0.046	0.022	0.014	0.008	0.008
SERCA(3)	0.018	0.007	0.008	0.005	0.004

**Table 1. Std.Dev. of  $\bar{p}(\tau)$  on the *wine* data set. The first row denotes the positive constraint ratio  $pr$ , below which are Std. Dev. of three methods' results respectively.**

test algorithm's performance at different scales. From the results in Fig.1 and Table 1, several effects can clearly be seen:

- As being expected, the performance of RCA and ISERCA is generally better than PCA, for the incorporation of side-information during the feature extraction procedure.
- Because ISERCA also utilizes the information implied by negative constraints and unlabeled data in an iterative way, it is generally more efficient and robust than RCA, especially when the positive constraint ratio is relatively low.

Furthermore, we did a more realistic experiment to test ISERCA's performance over data set with higher dimension. We selected 45 people from the Cohn-Kanade Facial Expression Database[6] and picked up their three different facial expressions - calmness, surprise and laugh - to form a data set. Each image is of  $16 \times 16$ . Some examples are illustrated below:

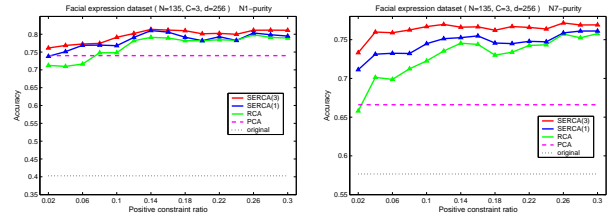


**Figure 2. Some example images from Cohn-Kanade Facial Expression Database.**

Facial expression is the underlying label. The experiment results are presented in Fig 3. We can see that the two effects stated above are also true in this case.

## 5 Conclusions

In this paper, we present a general iterative framework for relevant linear feature extraction with side-information and unlabeled data. Meanwhile the ISERCA algorithm is proposed to enhance RCA under this iterative framework. The contribution could be summarized into two aspects. First, information implied by both labeled and unlabeled



**Figure 3.  $\bar{p}(1)$  and  $\bar{p}(\tau)$  computed in five spaces over the reduced facial expression data set. The horizon axis denotes the positive constraint ratio, and the vertical axis denotes the mean neighbor purity. Negative constraint ratio is fixed as 0.1.**

data is sufficiently considered in this way. Second, an iterative framework is proposed to enhance algorithms' performance gradually. Experimental results show that the ISERCA algorithm is more efficient and robust than RCA, especially when the positive constraint ratio is relatively low - the regular case in many practical applications. Furthermore the proposed iterative framework can be applied to other linear and nonlinear methods related with side-information, such as Xing et. al.'s method in [8] and Kiri et. al.'s constrained K-means clustering in [7].

## References

- [1] N. Sental, T. Hertz, D. Weinshall, M. Pavel, Adjustment Learning and Relevant Component Analysis, in Proc. Of the 7th European Conference on Computer Vision, pp. 776-90, 2002.
- [2] A. Bar-Hillel, T. Hertz, N. Sental, D. Weinshall, Learning Distance Functions Using Equivalence Relations, in Proc. of 20th International Conference on Machine Learning, pp.11-18, 2003.
- [3] T. Hertz, A. Bar-Hillel, N. Sental, D. Weinshall, Learning Distance Functions with Product Space Boosting, Hebrew University, Leibniz Center for Research in Computer Science, TR 2003-35, 2003.
- [4] N. Sental, A. Bar-Hillel, T. Hertz, D. Weinshall, Computing Gaussian Mixture Models with EM Using Side-Information, in workshop "The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining"(with ICML-2003).
- [5] F. Alche-Bue, Y. Grandvalet, C. Ambroise, Semi-Supervised Marginboost, in Proc. of Nips, 2001.
- [6] T. Kanade, J. Cohn, Y. Tian, Comprehensive Database for Facial Expression Analysis, in Proc. of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00), March, 2000, pp. 46 - 53.
- [7] K. Wagstaff, C. Cardie, Constrained K-Means Clustering with Background Knowledge, in Proc. of 18th International Conference on Machine Learning, pp.577-584.
- [8] E. Xing, A. Ng, M. Jordan, S. Russell, Distance Metric Learning, with Application to Clustering with side-Information, Advances in Neural Information Processing Systems, MIT Press.