ProCRC: A Probabilistic Collaborative Representation based Classifier

Lei ZHANG
Dept. of Computing
Hong Kong Polytechnic University
http://www.comp.polyu.edu.hk/~cslzhang

Outline

- Related Work
  - Distance based classifier
  - NNC, NSC, SRC and CRC
- ProCRC
  - Probabilistic collaborative subspace
  - Probabilistic representation of samples outside the subspace
  - Probability to each class-specific subspace
  - The ProCRC model
- Experimental Results
- Discussions and Conclusion

Distance-based classifier

- Problem:
  - Training data matrix: $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_K]$.
  - $\boldsymbol{X}_k$, $k = 1,2, \ldots, K$ is the sample matrix of class k.
  - How to classify a given test sample $\boldsymbol{y}$?

- Classification by distance
  - Define the distance from $\boldsymbol{y}$ to class k:

$$d_k = d(\boldsymbol{y}, \boldsymbol{X}_k)$$

  - Classification rule:

$$Label(\boldsymbol{y}) = \text{argmin}_k\{d_k\}$$

  - **Question**: how to define d_k?

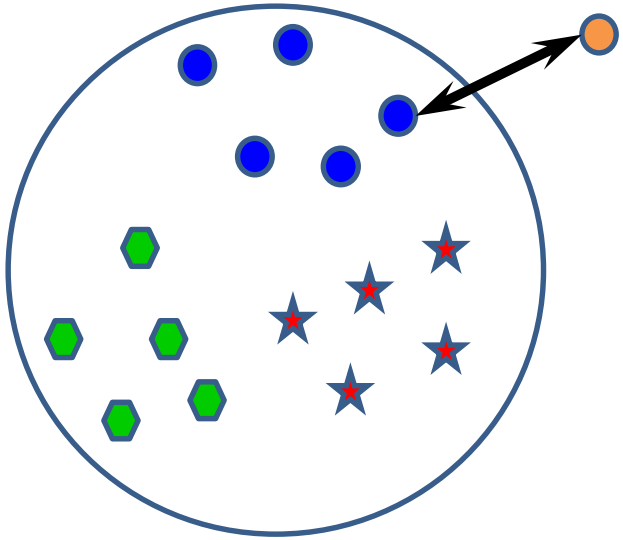
Nearest neighbor classifier (NNC)

- NNC

$$d(\boldsymbol{y}, \boldsymbol{X}_k) = \min_i d(\boldsymbol{y}, \boldsymbol{x}_{k,i}) = \min_i \left\| \boldsymbol{y} - \boldsymbol{x}_{k,i} \right\|_2$$

- The coefficients can written as

$$\boldsymbol{\alpha}_{k,l} = \begin{cases} 1 & \text{if } l = \operatorname{argmin}_i \left\| \boldsymbol{y} - \boldsymbol{x}_{k,i} \right\|_2^2 \\ 0 & \text{else} \end{cases}$$

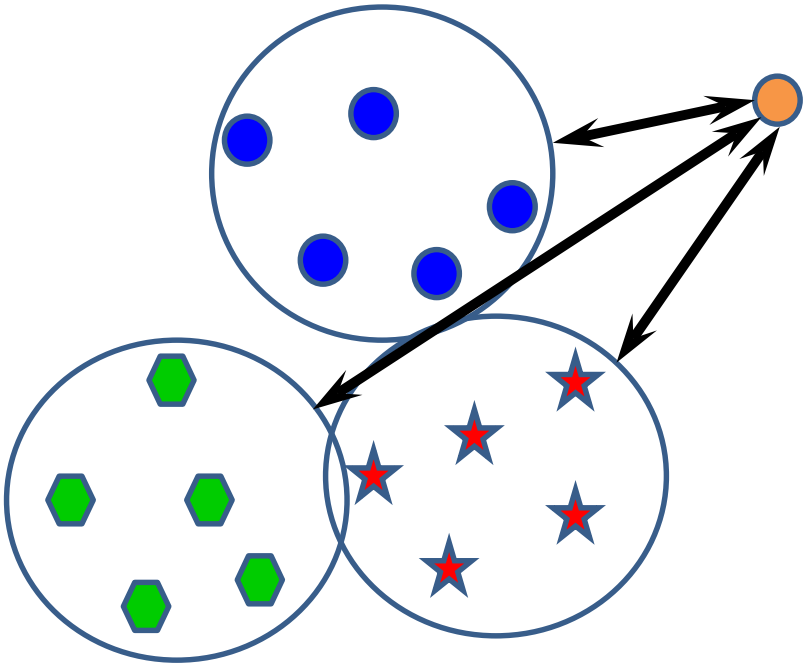


Nearest subspace classifier (NSC)

- NSC

$$\hat{\boldsymbol{\alpha}}_k = \text{argmin}_{\boldsymbol{\alpha}_k} \|\boldsymbol{y} - \boldsymbol{X}_k \boldsymbol{\alpha}_k\|_2^2$$

- NSC uses all the training samples within each class $\boldsymbol{X}_k$ to compute the distance from query sample $\boldsymbol{y}$ to class k.

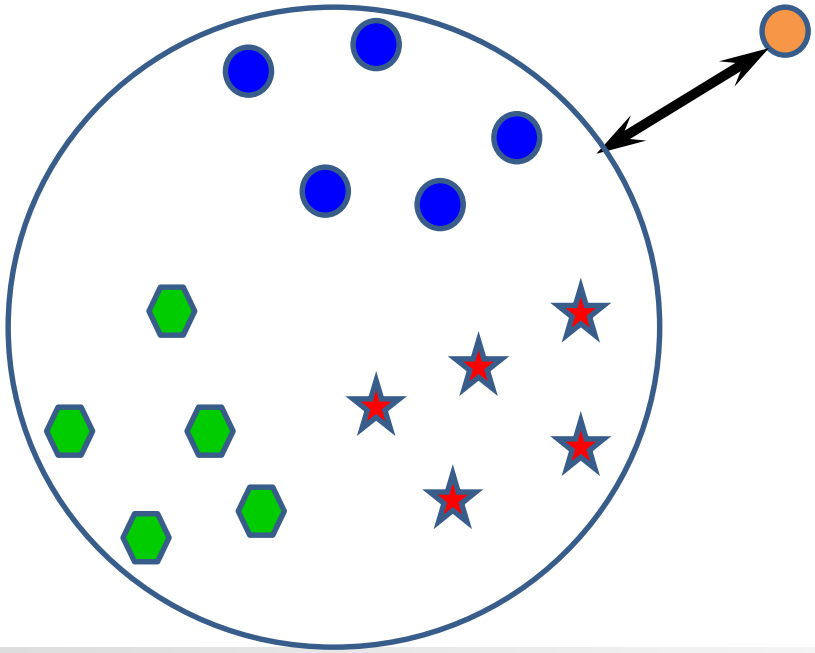


Sparse representation based classification (SRC, Wright, et al., PAMI09)

- SRC

$$\hat{\boldsymbol{\alpha}} = \text{argmin}_{\boldsymbol{\alpha}}\{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1\}$$

- SRC uses the training samples across all classes $\boldsymbol{X}$ to compute the coefficients of all classes simultaneously.

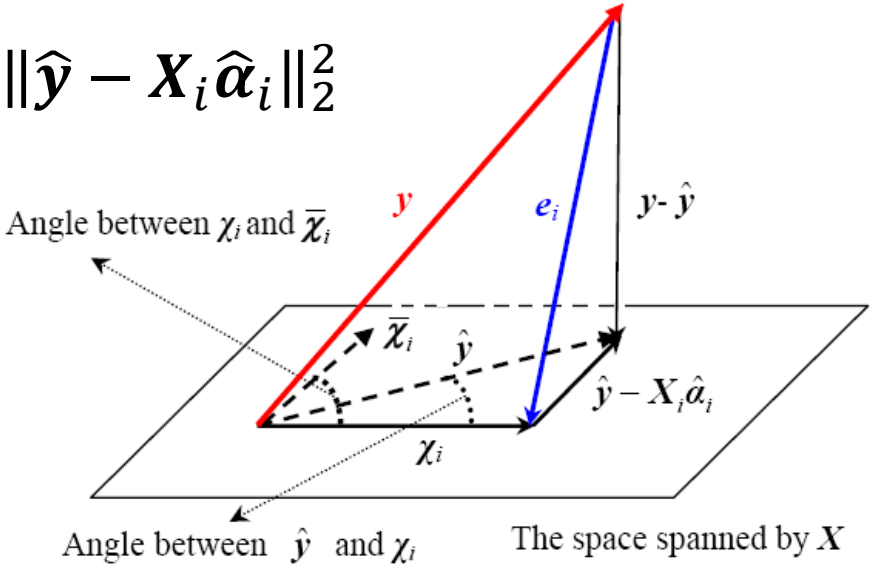


Collaborative representation based classification (CRC, Zhang, et al., ICCV2011)

$$(\hat{\boldsymbol{\alpha}}) = \text{argmin}_{\boldsymbol{\alpha}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}\|_2^2 \Rightarrow \hat{\boldsymbol{y}} = \sum_i \boldsymbol{X}_i \hat{\boldsymbol{\alpha}}_i$$

$$e_i = \|\boldsymbol{y} - \boldsymbol{X}_i \hat{\boldsymbol{\alpha}}_i\|_2^2 = \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2^2 + \|\hat{\boldsymbol{y}} - \boldsymbol{X}_i \hat{\boldsymbol{\alpha}}_i\|_2^2$$

$$e_i^* = \frac{\sin^2(\hat{\boldsymbol{y}}, \boldsymbol{\chi}_i)\|\hat{\boldsymbol{y}}\|_2^2}{\sin^2(\boldsymbol{\chi}_i, \overline{\boldsymbol{\chi}_i})}$$



Only $e_i^* = \|\hat{\boldsymbol{y}} - \boldsymbol{X}_i \hat{\boldsymbol{\alpha}}_i\|_2^2$ works for classification

Collaborative representation model

$$\min_{\alpha} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}\|_{l_q} + \lambda\|\boldsymbol{\alpha}\|_{l_q} \quad p, q = 1 \; or \; 2$$

q=2,p=1, Sparse Representation based Classification (S-**SRC**)

q=2,p=2, Collaborative Representation based Classification with regularized least square (CRC_RLS)

q=1,p=1, Robust Sparse Representation based Classification (R-**SRC**)

q=1,p=2, Robust Collaborative Representation based Classification (R-CRC)

CRC_RLS has a closed-form solution; others have iterative solutions.

Why SRC/CRC works?

- SRC/CRC represents the query image by gallery images from all classes. However, it uses the representation residual by each class for classification.
- What kind of classifier SRC/CRC is?
- Why SRC/CRC works?

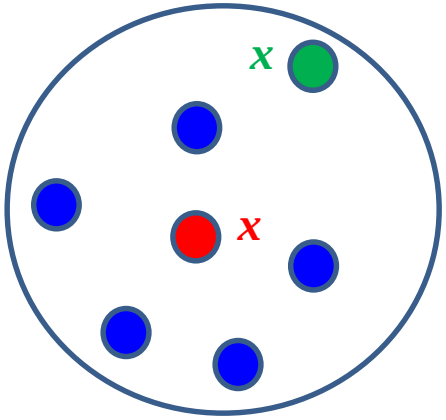
S. Cai, L. Zhang, W. Zuo, and X. Feng, "A Probabilistic Collaborative Representation based Approach for Pattern Classification," CVPR 2016.

Probabilistic collaborative subspace

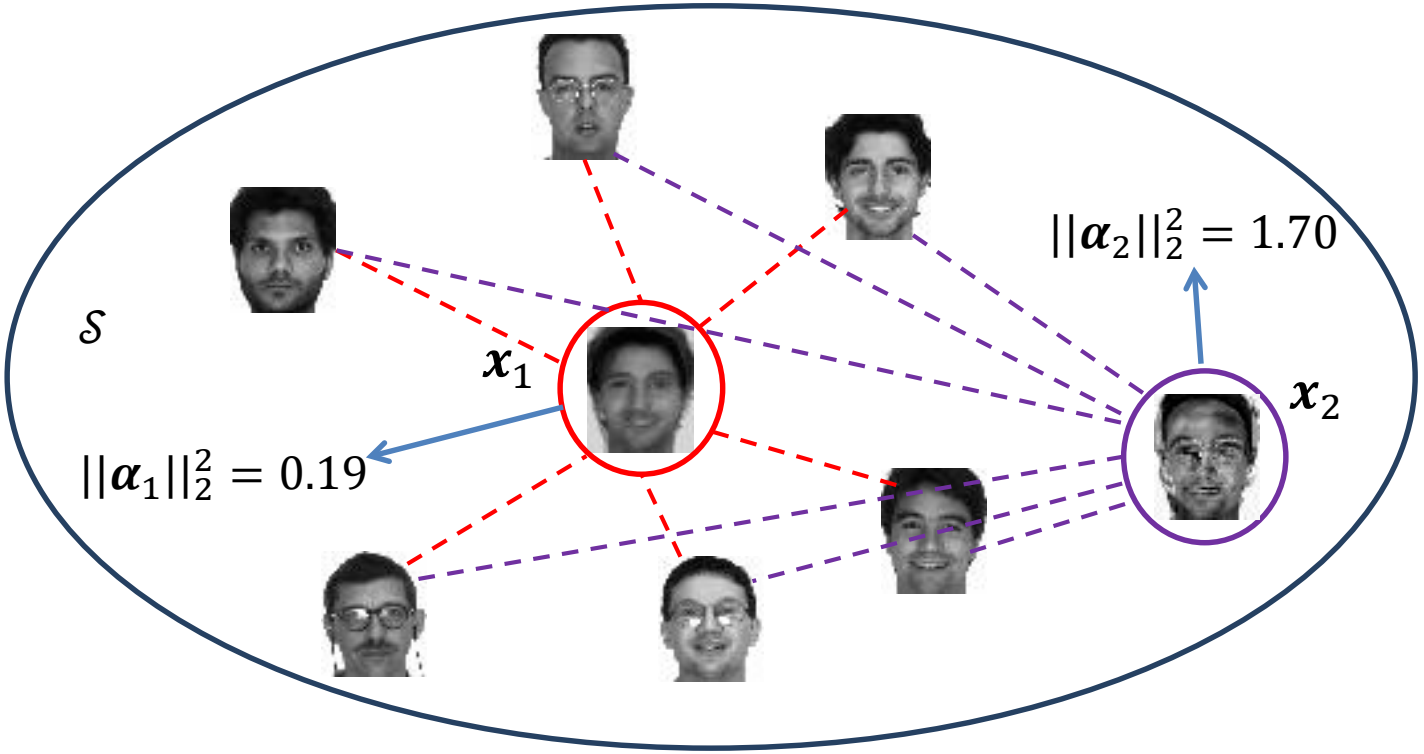
- Training samples from K classes:

$$\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_K].$$

- $l_{\boldsymbol{X}}$: the label set of all classes in $\boldsymbol{X}$
- $\mathcal{S}$: the subspace spanned by $\boldsymbol{X}$
  - Each data point $\boldsymbol{x}$ within $\mathcal{S}$ can be written as: $\boldsymbol{x} = \boldsymbol{X}\boldsymbol{\alpha}$.
- How can we characterize the confidence that $l(\boldsymbol{x}) \in l_{\boldsymbol{X}}$?



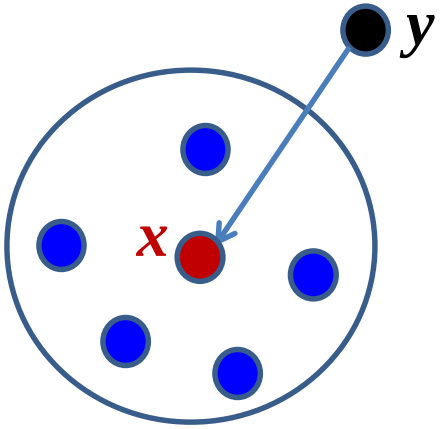
Probabilistic collaborative subspace



- We define the probability that $l(\boldsymbol{x}) \in l_{\boldsymbol{X}}$ as :

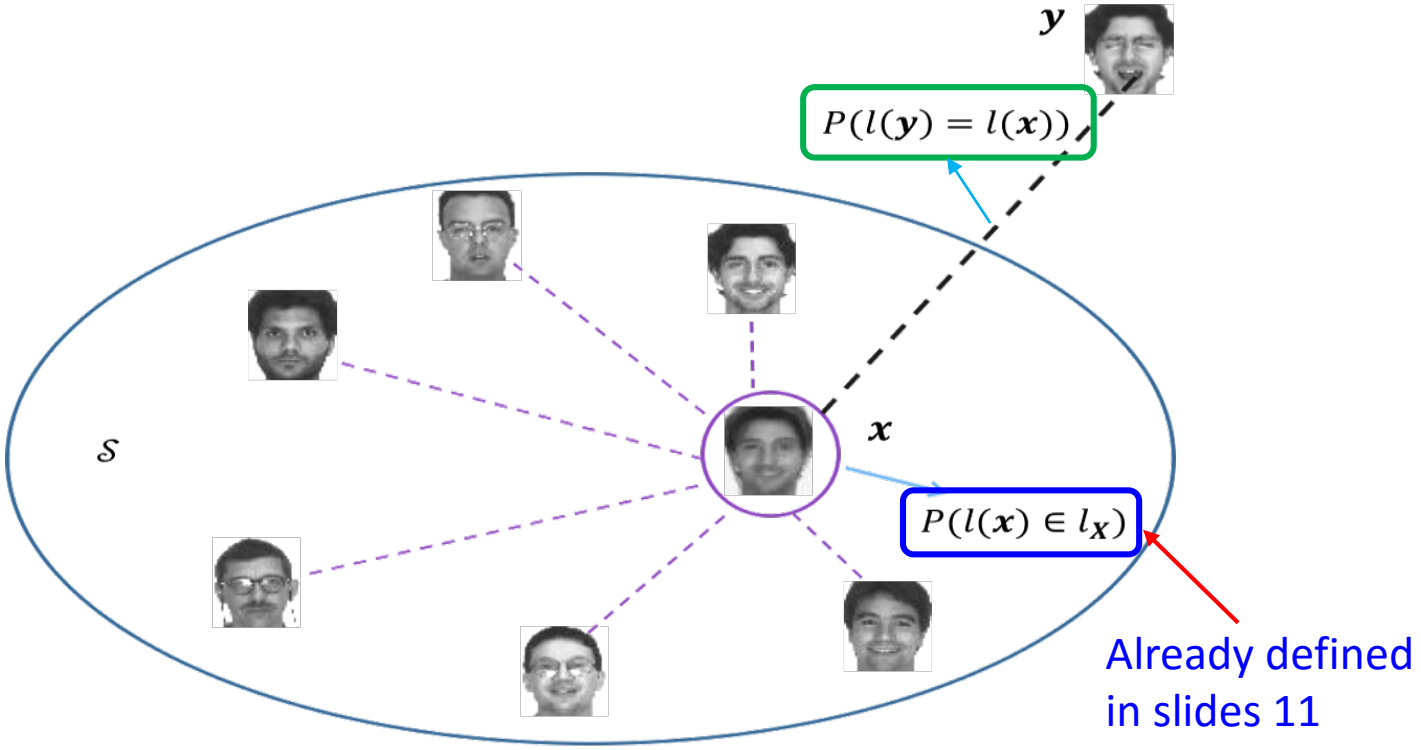
$$\boldsymbol{P(l(x) \in l_X) \propto \exp(-c\|\alpha\|_2^2)}$$

Outside the collaborative subspace



- The query sample $\boldsymbol{y}$ usually lies outside the collaborative subspace $\mathcal{S}$. What is the probability that $\boldsymbol{y}$ belongs to $l_{\boldsymbol{X}}$?
- It can be determined by two factors:
  - Given $\boldsymbol{x} = \boldsymbol{X\alpha}$, how likely $\boldsymbol{y}$ has the same class label as $\boldsymbol{x}$?
  - What is the probability that $\boldsymbol{x}$ belongs to $l_{\boldsymbol{X}}$?

Outside the collaborative subspace



$$P(l(\boldsymbol{y}) \in l_X) = P(l(\boldsymbol{y}) = l(\boldsymbol{x}) | l(\boldsymbol{x}) \in l_X) \cdot P(l(\boldsymbol{x}) \in l_X)$$

Outside the collaborative subspace

- We adopt Gaussian kernel to measure the label-consistent probability:

$$P(l(\boldsymbol{y}) = l(\boldsymbol{x}) | l(\boldsymbol{x}) \in l_{\boldsymbol{X}}) \propto \exp(-\kappa \|\boldsymbol{y} - \boldsymbol{x}\|_2^2)$$

- Then we have

$$P(l(\boldsymbol{y}) \in l_{\boldsymbol{X}}) \propto \exp(-\kappa \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}\|_2^2 + c\|\boldsymbol{\alpha}\|_2^2)$$

Probability to each class-specific subspace

- $\boldsymbol{x} = \boldsymbol{X}\boldsymbol{\alpha} = \sum_{k=1}^{K} \boldsymbol{X}_k \boldsymbol{\alpha}_k = \sum_{k=1}^{K} \boldsymbol{x}_k$
- The probability that $\boldsymbol{x}$ has the same class label as $\boldsymbol{x}_k$:

$$P(l(\boldsymbol{x}) = k | l(\boldsymbol{x}) \in l_{\boldsymbol{X}}) \propto \exp(-\delta \|\boldsymbol{x} - \boldsymbol{X}_k \boldsymbol{\alpha}_k\|_2^2)$$

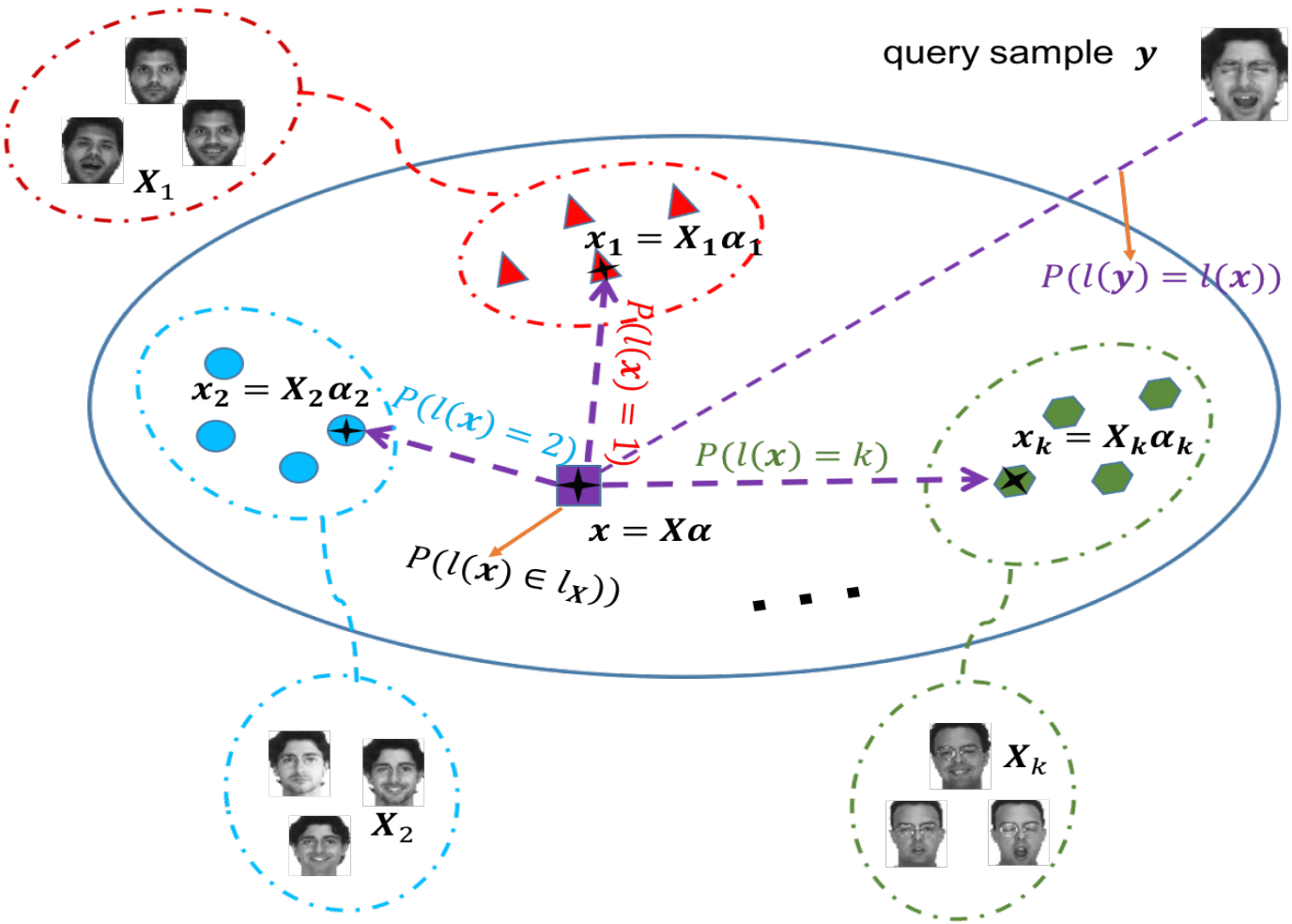
- For a query sample y outside the space $\mathcal{S}$:

$$P(l(\boldsymbol{y}) = k) = P(l(\boldsymbol{y}) = l(\boldsymbol{x}) | l(\boldsymbol{x}) = k) \cdot P(l(\boldsymbol{x}) = k | l(\boldsymbol{x}) \in l_{\boldsymbol{X}}) \cdot P(l(\boldsymbol{x}) \in l_{\boldsymbol{X}})$$

- Since $P(l(\boldsymbol{y}) = l(\boldsymbol{x}) | l(\boldsymbol{x}) = k) = P(l(\boldsymbol{y}) = l(\boldsymbol{x}) | l(\boldsymbol{x}) \in l_{\boldsymbol{X}})$, we have

$$P(l(\boldsymbol{y}) = k) = P(l(\boldsymbol{y}) \in l_{\boldsymbol{X}}) \cdot P(l(\boldsymbol{x}) = k | l(\boldsymbol{x}) \in l_{\boldsymbol{X}})$$

Probability to each class-specific subspace



$$P(l(\boldsymbol{y}) = k) \propto \exp(-(\|\boldsymbol{y} - \boldsymbol{X\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_2^2 + \gamma\|\boldsymbol{X\alpha} - \boldsymbol{X}_k\boldsymbol{\alpha}_k\|_2^2))$$

The ProCRC model

- Finding a common data point $\boldsymbol{x} = \boldsymbol{X}\boldsymbol{\alpha}$, i.e., $\boldsymbol{\alpha}$, that maximizes the joint probability:

$$\max P(l(\boldsymbol{y}) = 1, \dots, l(\boldsymbol{y}) = K) = \max \prod_{k=1}^{K} P(l(\boldsymbol{y}) = k)$$

$$\propto \max \exp(-(\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_2^2 + \frac{\gamma}{K}\sum_{k=1}^{K}\|\boldsymbol{X}\boldsymbol{\alpha} - \boldsymbol{X}_k\boldsymbol{\alpha}_k\|_2^2))$$

- Applying the log-operator:

$$\hat{\boldsymbol{\alpha}} = \text{argmin}_{\alpha}\{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_2^2 + \frac{\gamma}{K}\sum_{k=1}^{K}\|\boldsymbol{X}\boldsymbol{\alpha} - \boldsymbol{X}_k\boldsymbol{\alpha}_k\|_2^2\}$$

ProCRC: classification rule

- We use the marginal probability for classification :

$$P(l(\boldsymbol{y}) = k) \propto \exp\left(-\left(\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\alpha}}\|_2^2 + \lambda\|\hat{\boldsymbol{\alpha}}\|_2^2 + \frac{\gamma}{K}\|\boldsymbol{X}\hat{\boldsymbol{\alpha}} - \boldsymbol{X}_k\hat{\boldsymbol{\alpha}}_k\|_2^2\right)\right)$$

$$\propto \exp\left(-\|\boldsymbol{X}\hat{\boldsymbol{\alpha}} - \boldsymbol{X}_k\hat{\boldsymbol{\alpha}}_k\|_2^2\right)$$

- The classification rule is:

$$l(\boldsymbol{y}) = \operatorname{argmax}_k\{P(l(\boldsymbol{y}) = k)\} = \operatorname{argmin}_k\{\|\boldsymbol{X}\hat{\boldsymbol{\alpha}} - \boldsymbol{X}_k\hat{\boldsymbol{\alpha}}_k\|_2^2\}$$

The robust ProCRC model

- Using the Laplacian kernel to measure the label-consistent probability:

$$P(l(\boldsymbol{y}) = l(\boldsymbol{x}) | l(\boldsymbol{x}) \in l_{\boldsymbol{X}}) \propto \exp(-\kappa \|\boldsymbol{y} - \boldsymbol{x}\|_1)$$

- Robust ProCRC model (R-ProCRC):

$$\min_{\boldsymbol{\alpha}} \{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}\|_1 + \lambda \|\boldsymbol{\alpha}\|_2^2 + \frac{\gamma}{K} \sum_{k=1}^{K} \|\boldsymbol{X}\boldsymbol{\alpha} - \boldsymbol{X}_k \boldsymbol{\alpha}_k\|_2^2\}$$

Handwritten digit recognition: MNIST

Number of training samples per class	50	100	300	500
SVM	89.35	92.10	94.88	**95.93**
NSC	91.06	92.86	85.29	78.26
CRC	72.21	82.22	86.54	87.46
SRC	80.12	85.63	89.30	92.70
CROC	91.06	92.86	89.93	89.37
ProCRC	**92.16**	**94.56**	**95.58**	95.88

Handwritten digit recognition: USPS

Number of training samples per class	50	100	200	300
SVM	93.46	95.31	95.91	96.30
NSC	93.48	93.25	90.21	87.85
CRC	89.89	91.67	92.36	92.79
SRC	92.58	93.99	95.63	95.86
CROC	93.48	93.25	91.40	91.87
ProCRC	**93.84**	**95.62**	**96.03**	**96.43**

Robust face recognition (YaleB)

- Random corruption

Corruption ratio	10%	20%	40%	60%
R-SRC	97.49	95.60	90.19	76.85
R-ProCRC	**98.45**	**98.20**	**93.25**	**82.42**

- Block occlusion

Occlusion ratio	10%	20%	30%	40%
R-SRC	90.42	85.64	78.89	70.09
R-ProCRC	**98.12**	**92.62**	**86.42**	**77.16**

Robust face recognition (AR)

- Disguise

Disguise	Sunglasses	Scarf
R-SRC	69.17	69.50
R-ProCRC	**70.50**	**69.83**

Running time

- Intel Core (TM) i7-5930K 3.50 GHz CPU with 32 GB RAM
- Running time (second) of different methods on the MNIST dataset:

Method	NSC	CRC	SRC	CROC
Times (s)	0.0003	0.0005	0.22	0.0009
Method	**ProCRC**	**R-SRC**	**R-ProCRC**	
Times (s)	0.0005	3.57	1.81	

Performance with SIFT and CNN features

Classifier		Softmax	SVM	K-SVM	NSC	CRC	SRC	CROC	ProCRC
Standford 40	BOW-SIFT	21.1	24.0	26.3	22.1	24.6	24.2	24.5	**28.4**
	VGG19	77.2	79.0	79.8	74.7	78.2	78.7	79.1	**80.9**

40 human actions
9352 images

B. Yao, X. Jiang, A. Khosla, A.L. Lin, L.J. Guibas, and L. Fei-Fei. Human Action Recognition by Learning Bases of Action Attributes and Parts. In *ICCV* 2011.

Performance with SIFT and CNN features

Classifier		Softmax	SVM	K-SVM	NSC	CRC	SRC	CROC	ProCRC
CUB200-2011	BOW-SIFT	8.2	10.2	10.5	8.4	9.4	7.7	9.1	**9.9**
	VGG19	72.1	75.4	76.6	74.5	76.2	76.0	76.2	**78.3**

200 bird species
11,788 images

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Performance with SIFT and CNN features

Classifier		Softmax	SVM	K-SVM	NSC	CRC	SRC	CROC	ProCRC
Flower 102	BOW-SIFT	46.5	50.1	51.0	46.7	49.9	47.2	49.4	**51.2**
	VGG19	87.3	90.9	92.2	90.1	93.0	93.2	93.1	**94.8**

102 flower categories
8,189 images

M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *CVGIP* 2008.

Performance with SIFT and CNN features

Classifier		Softmax	SVM	K-SVM	NSC	CRC	SRC	CROC	ProCRC
Caltech 256 (30)	BOW-SIFT	21.1	24.0	26.3	22.1	24.6	24.2	24.5	**28.4**
	VGG19	77.2	79.0	79.8	74.7	78.2	78.7	79.1	**80.9**

256 object categories
30,608 images

G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. 2007.

Comparison with state-of-the-arts

Dataset	Split	Methods & Accuracies (%)					
Standford 40	fixed	ProCRC	ASPD	SMP	CF	SparBase	EPM
		80.9	75.4	53.0	51.9	45.7	42.2
CUB200-2011	fixed	ProCRC	NAC	PN-CNN	FV-CNN	POOF	
		78.3	**81.0**	75.7	66.7	56.9	
Flower 102	fixed	ProCRC	NAC	OverFeat	GMP	DAS	BiCos
		94.8	**95.3**	86.8	84.6	80.7	79.4
Caltech-256	random	ProCRC	NAC	VGG19	CNN-S	ZF	M-HMP
	15	**80.2**	-	-	-	65.7	42.7
	30	**83.3**	-	-	-	70.6	50.7
	45	**84.9**	-	-	-	72.7	54.8
	60	**86.1**	84.1	85.1	77.6	74.2	58.0

Discussions

- **Scalability with many samples per class**
  - use simple dictionary learning (DL) model to compact the training set
- **Large number of classes**
  - cluster all classes into a tree-like structure with super-classes and perform level-wise ProCRC
- **End-to-end learning with deep architectures**
  - Joint learning with CNN features
  - E.g., DPL-CNN (CVPR16)

Conclusions

- ProCRC provides a good probabilistic interpretation of collaborative representation based classifiers (NSC, SRC and CRC).
- ProCRC achieves higher classification accuracy than the competing classifiers in most experiments.
- By introducing the simple dictionary learning pre-processing stage, ProCRC is still a competitive and efficient classifier on larger-scale datasets .

Thanks for your attention!