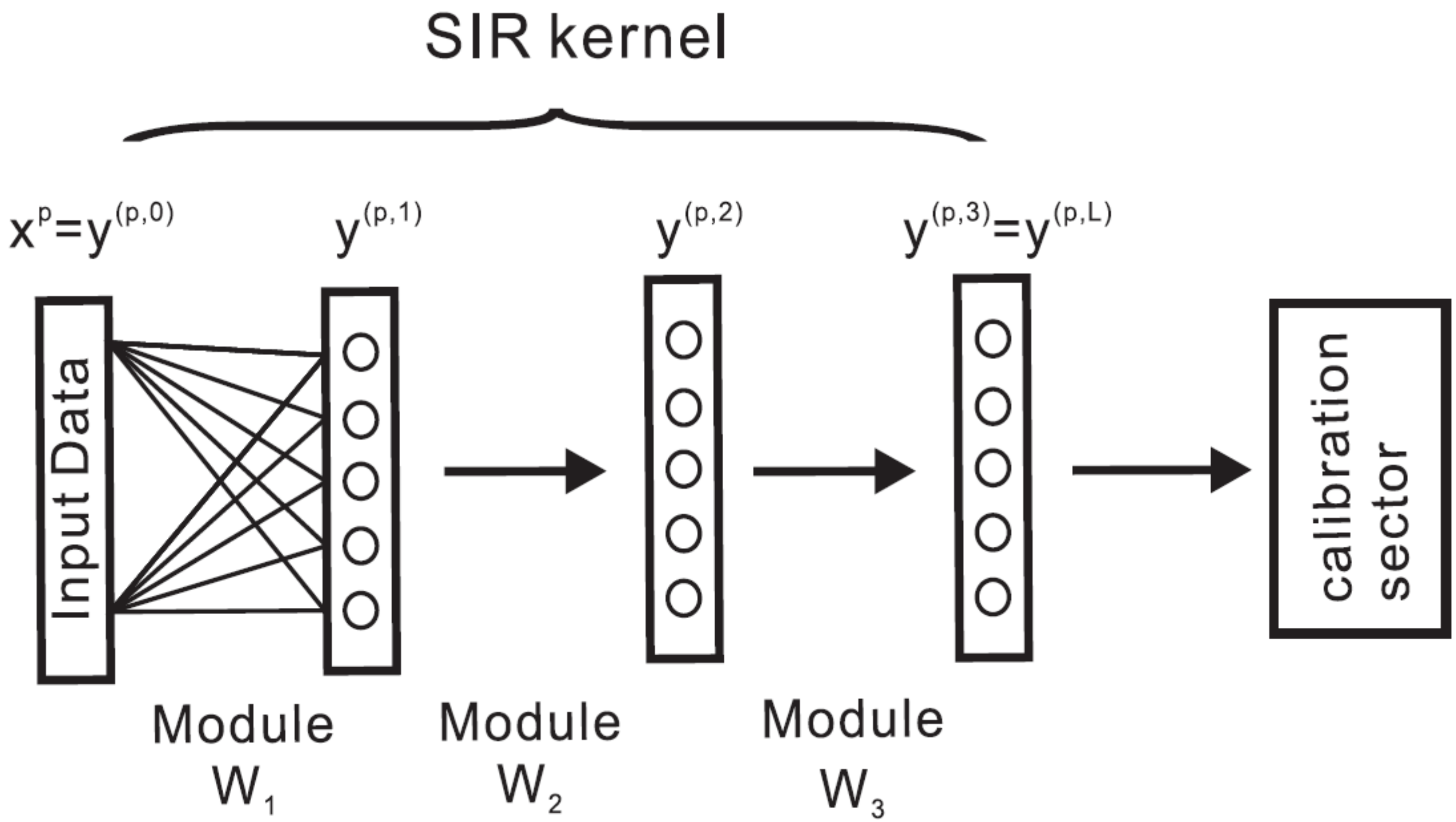
Implementation of the MLP Kernel

Cheng-Yuan Liou* and Wei-Chen Cheng
Department of Computer Science and Information Engineering
National Taiwan University
Republic of China
*cyliou@csie.ntu.edu.tw

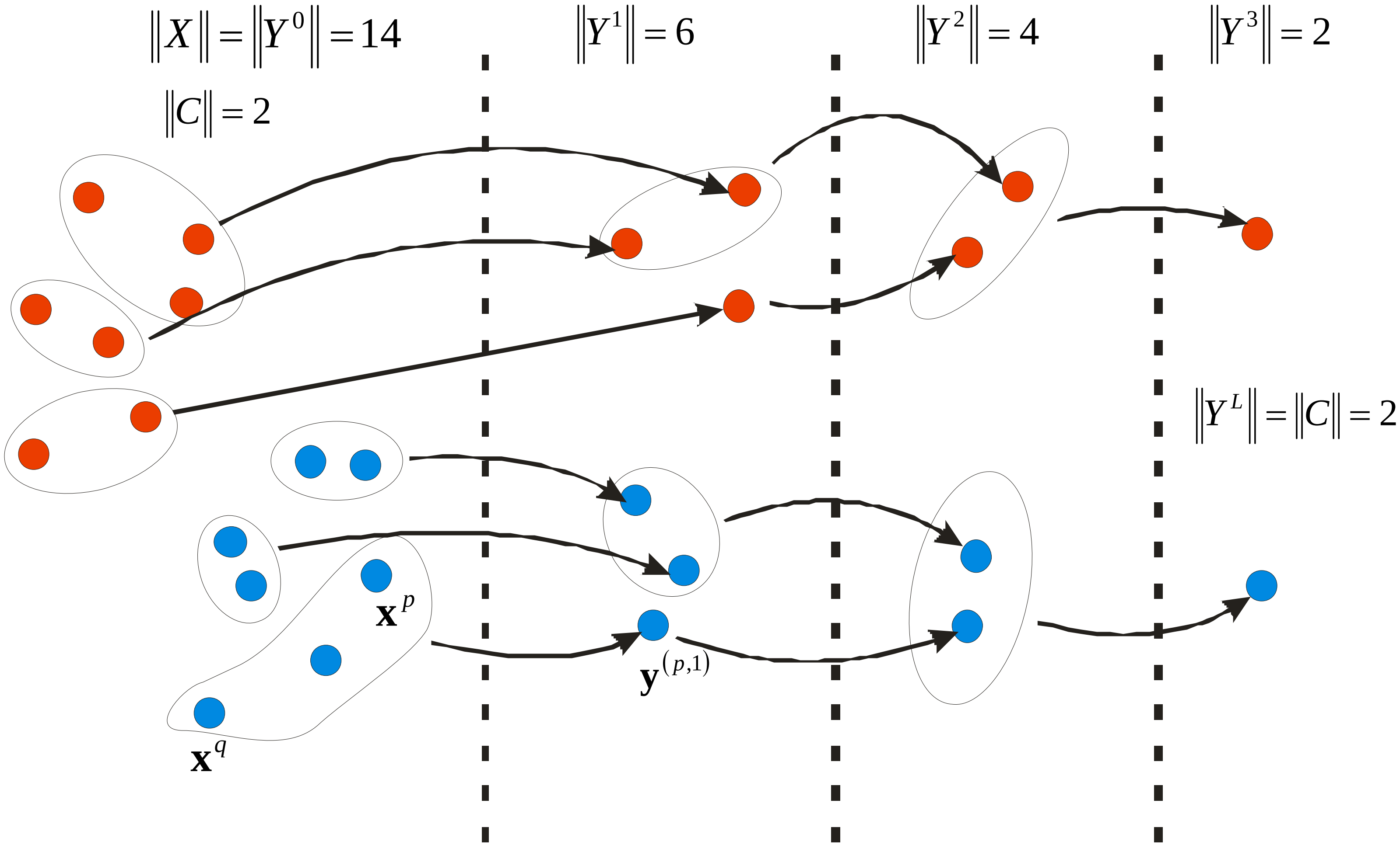
25th, Nov., 2008
17:40-19:00
Auckland


SIR kernel
x^p=y^(p,0)
y^(p,1)
y^(p,2)
y^(p,3)=y^(p,L)
Input Data
calibration sector
Module W1
Module W2
Module W3

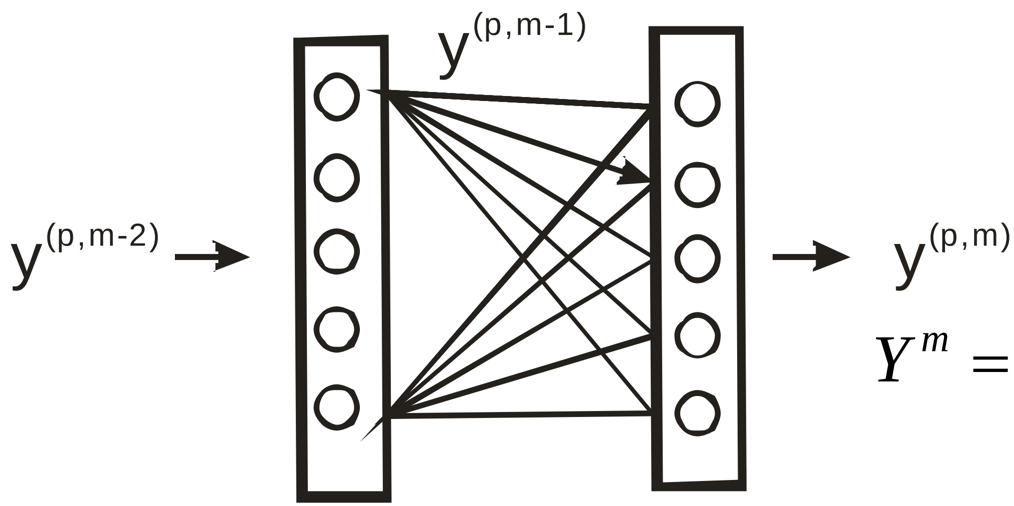
Related works

Year	People	Contribution
1992	Boser	Support vector machine
1994	Liou, Yu	ICONIP, Weight design, upper bound $n_m < \left\lceil \frac{\left\| Y^{m-1} \right\|}{n_{m-1}} \right\rceil$
1995	Liou, Yu	ICNN, Perth, AIR
2000	Liou, Chen, Huang	ICS, SIR
2007	Liou, Cheng	ICONIP

$\|Y^{m-1}\| << \|Y^m\|$, many-to-one mapping



Liou and Yu, 1995, ICNN, Perth



$$Y^m = \left\{ \mathbf{y}^{(p,m)} \right\}$$

m-1 m^th layer

W_m by design, $n_m < \left\lceil \frac{\|Y^{m-1}\|}{n_{m-1}} \right\rceil$ and $\|Y^L\| = \|C\|$ guaranteed

(Liou and Yu, 1994, ICONIP, Seoul)

W_m by training, SIR

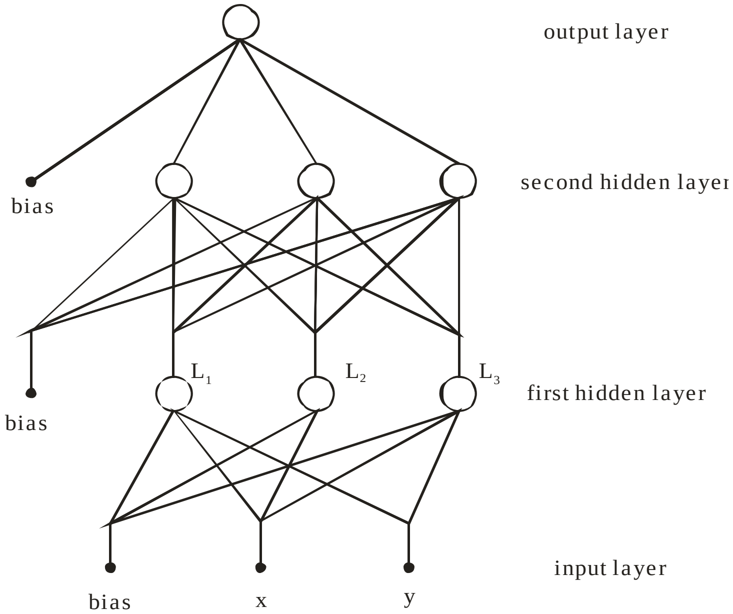
(Liou, Chen, Huang, 2000, ICS)

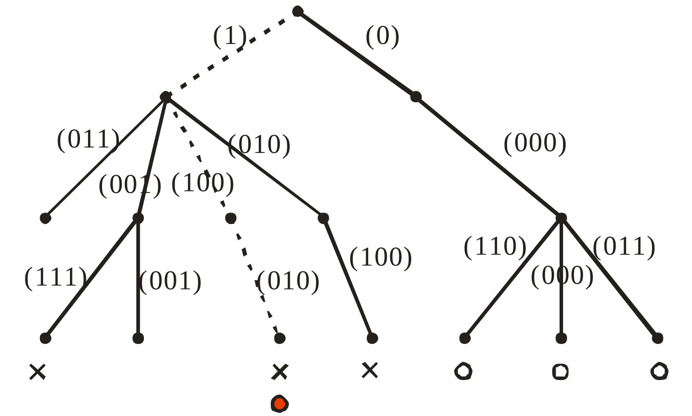
(Liou, Cheng, 2007, ICONIP)

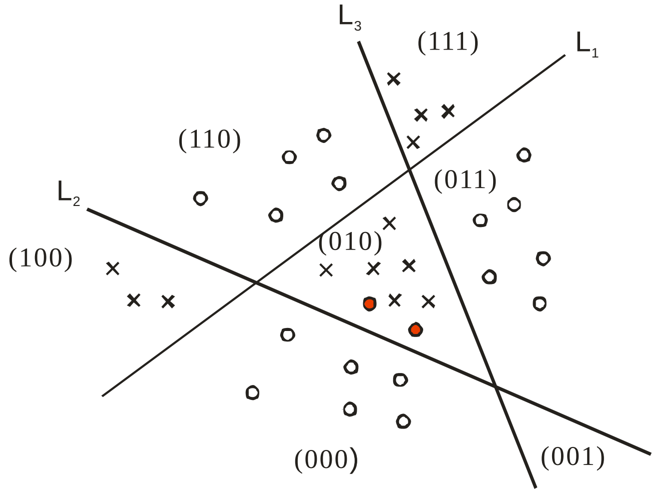
The reason why the training is layer after layer independently.

[Liou and Yu, 1995, ICNN, Perth]

ICNN, 1995, Perth, AIR tree







Conclusions of AIR, 1995

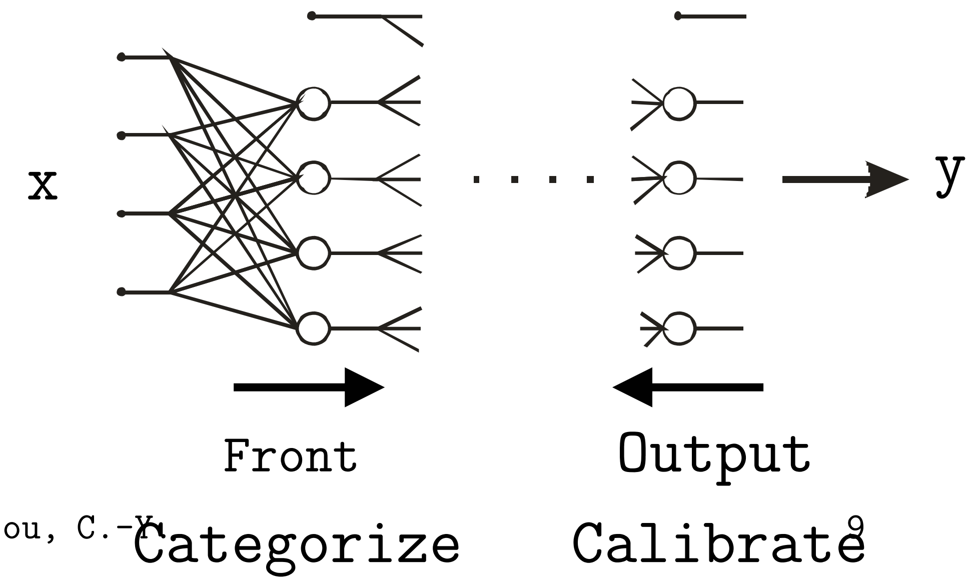
- BP can not correct the latent error neurons by adjusting their succeeding layers.
- AIR tree can trace the errors in a latent layer that near the front input layer.
- The front layers must send right signals to their succeeding layers.
- The front layer must be trained layer after layer in order to get right signals.
- Split the function of supervised BP, categorization and calibration.
- Reduced number of representations. $\|Y^{m-1}\| << \|Y^{m}\|$

AIR, 1995

- Supervised BP
- Identified the function of MLP
  - Classification =

Categorization + Calibration

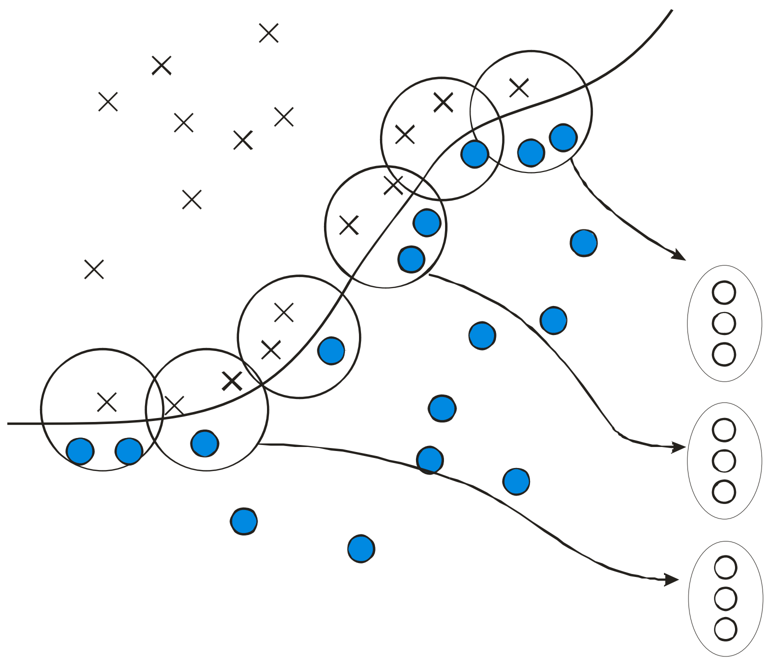
Differences of classes   class labels



Weight design
ICONIP, 1994, Seoul

- Weight design for each layer
- Number of neurons (E.B. Baum 1988)
  - Upper bound $\left\lceil \frac{P}{D} \right\rceil$ for first hidden layer
  - $n_m < \left\lceil \frac{\|Y^{m-1}\|}{n_{m-1}} \right\rceil$ for hidden layers
  - $\|Y^L\| = \|C\|$ the number of classes, guaranteed

Continuous border



divide and conquer

1st hidden layer

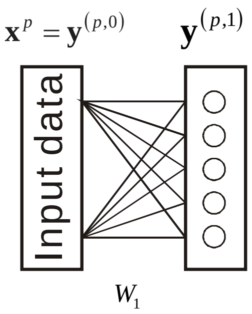
Devise training for layers SIR, ICS 2000

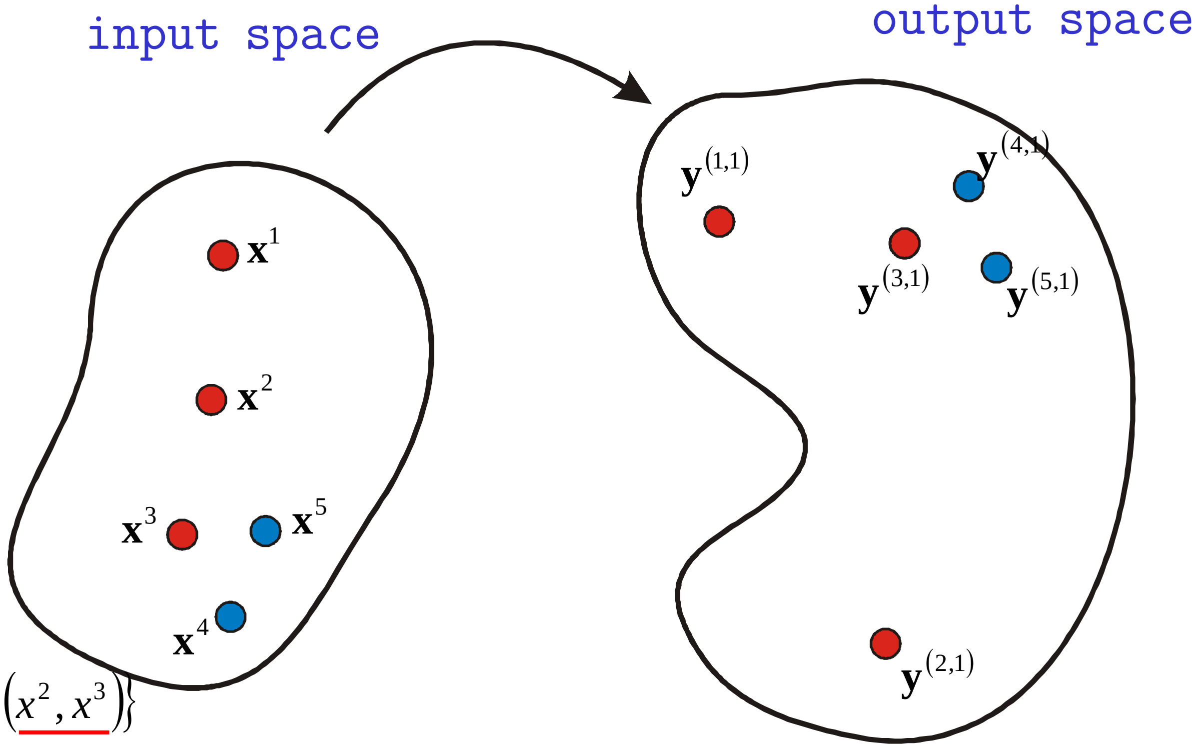
- Categorization sector
- Using differences between classes implicitly

$$E^{rep} = \frac{-1}{2}\left\| \mathbf{y}^{(p,m)} - \mathbf{y}^{(q,m)} \right\| \quad \text{Inter-class}$$

$$E^{att} = \frac{1}{2}\left\| \mathbf{y}^{(p,m)} - \mathbf{y}^{(q,m)} \right\| \quad \text{Intra-class}$$

SIR kernel





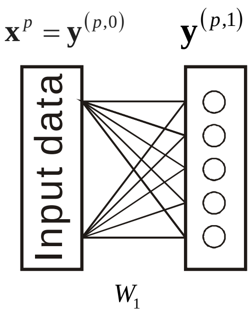
$U_1 = \{(x^1, x^2), (x^1, x^3), (x^2, x^3)\}$

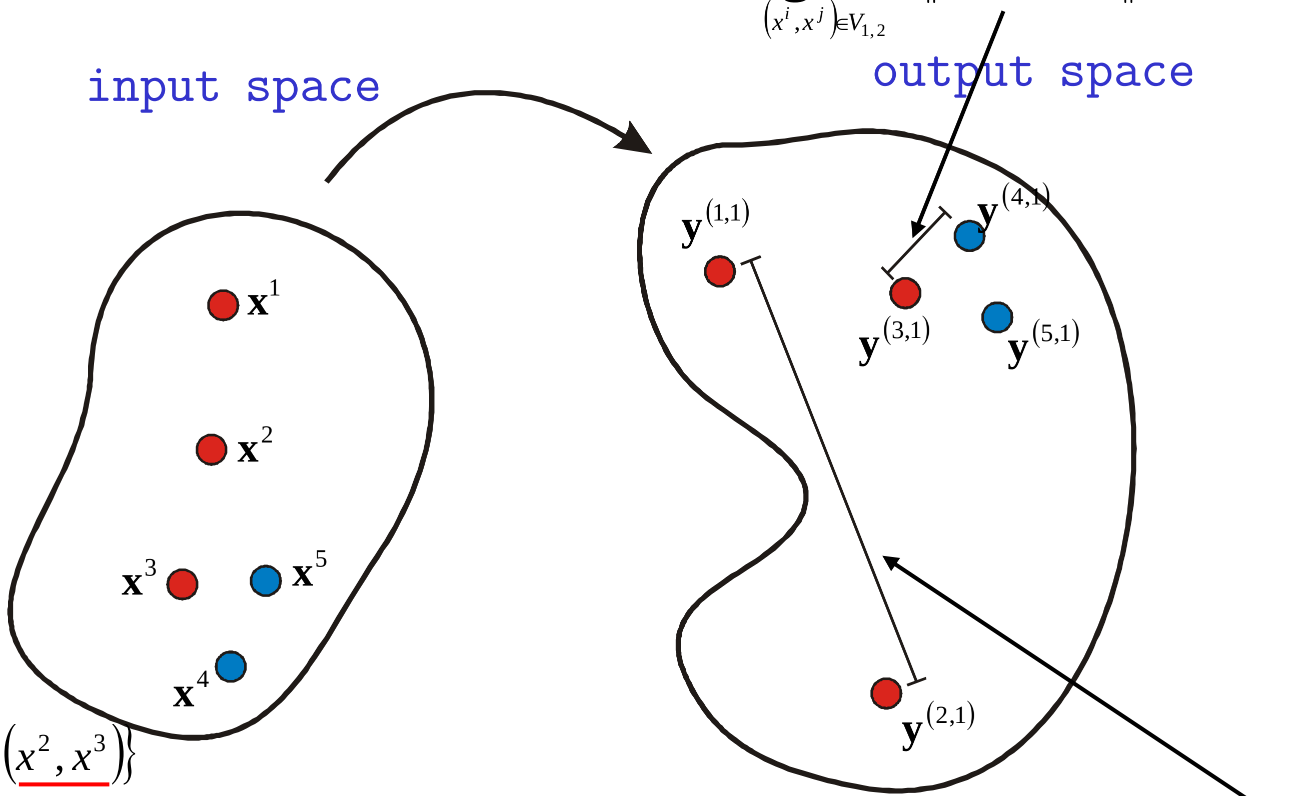
$U_2 = \{(x^4, x^5)\}$

$V_{1,2} = \{(x^1, x^4), (x^1, x^5), (x^2, x^4), (x^2, x^5), (x^3, x^4), (x^3, x^5)\}$: inter-class pattern pairs

SIR kernel

$$\left(x^3, x^4\right) = \underset{\left(x^i, x^j\right) \in V_{1,2}}{\arg\min} \left\| \mathbf{y}^{(i,1)} - \mathbf{y}^{(j,1)} \right\|$$





$$U_1 = \left\{\left(x^1, x^2\right), \left(x^1, x^3\right), \left(x^2, x^3\right)\right\}$$

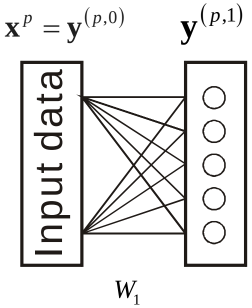
$$U_2 = \left\{\left(x^4, x^5\right)\right\}$$

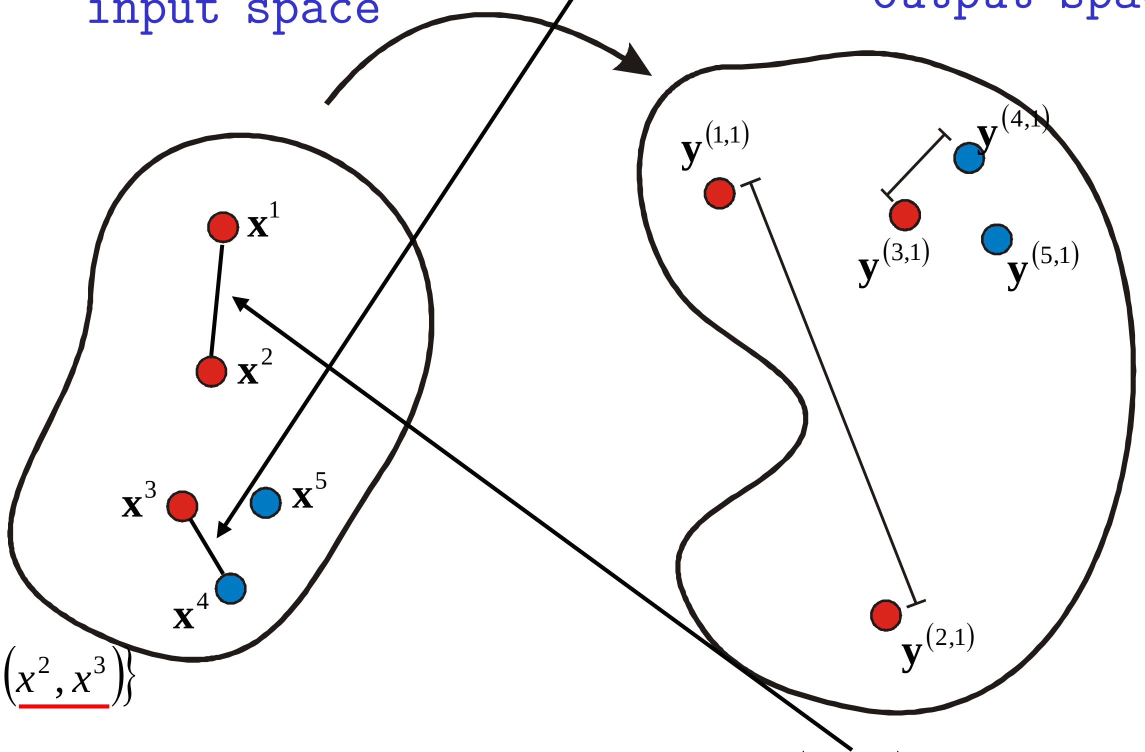
$$V_{1,2} = \left\{\left(x^1, x^4\right), \left(x^1, x^5\right), \left(x^2, x^4\right), \left(x^2, x^5\right), \left(x^3, x^4\right), \left(x^3, x^5\right)\right\}$$

$$\left(x^1, x^2\right) = \underset{\left(x^p, x^q\right) \in U_1 \text{ or } \left(x^p, x^q\right) \in U_2}{\arg\max} \left\| \mathbf{y}^{(p,1)} - \mathbf{y}^{(q,1)} \right\|$$

SIR kernel

$$\left(x^3, x^4\right) = \underset{\left(x^i, x^j\right) \in V_{1,2}}{\arg\min} \left\| \mathbf{y}^{(i,1)} - \mathbf{y}^{(j,1)} \right\|$$





$$U_1 = \left\{ \left(x^1, x^2\right), \left(x^1, x^3\right), \left(x^2, x^3\right) \right\}$$

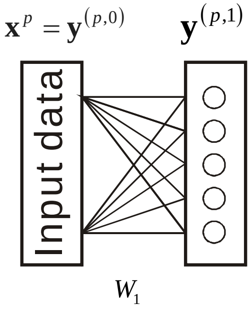
$$U_2 = \left\{ \left(x^4, x^5\right) \right\}$$

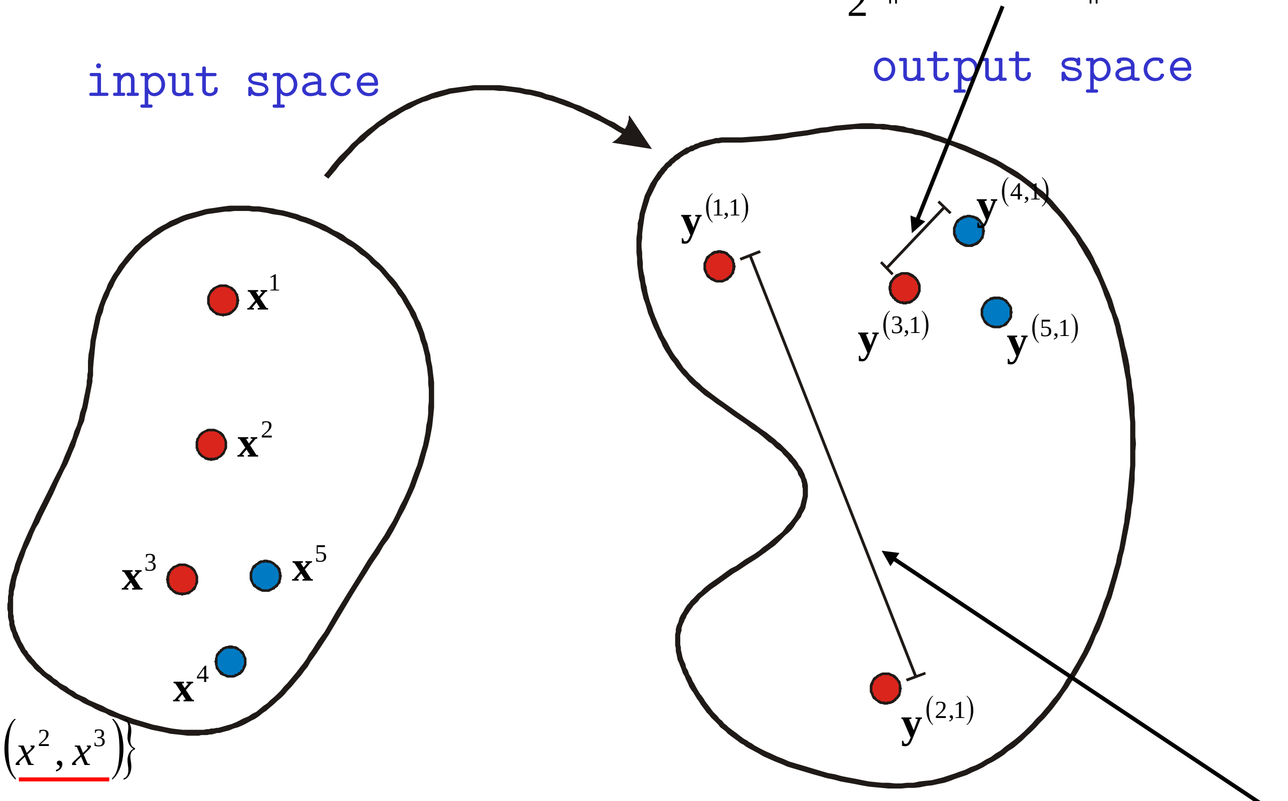
$$V_{1,2} = \left\{ \left(x^1, x^4\right), \left(x^1, x^5\right), \left(x^2, x^4\right), \left(x^2, x^5\right), \left(x^3, x^4\right), \left(x^3, x^5\right) \right\}$$

$$\left(x^1, x^2\right) = \underset{\left(x^p, x^q\right) \in U_1 \text{ or } \left(x^p, x^q\right) \in U_2}{\arg\max} \left\| \mathbf{y}^{(p,1)} - \mathbf{y}^{(q,1)} \right\|$$

SIR kernel

$$E^{rep}\left(\mathbf{x}^3,\mathbf{x}^4\right)=\frac{-1}{2}\left\|\mathbf{y}^{(3,1)}-\mathbf{y}^{(4,1)}\right\|$$





$$E^{att}\left(\mathbf{x}^1,\mathbf{x}^2\right)=\frac{1}{2}\left\|\mathbf{y}^{(1,1)}-\mathbf{y}^{(2,1)}\right\|$$

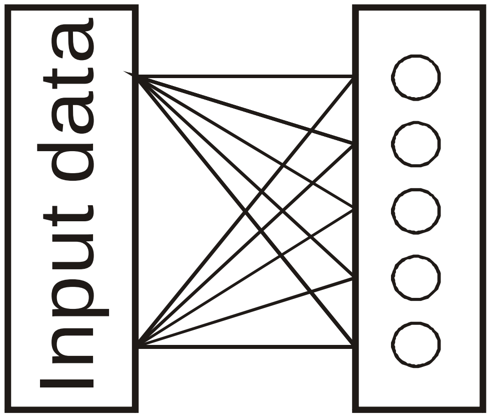
$$U_1=\left\{\left(x^1,x^2\right),\left(x^1,x^3\right),\left(x^2,x^3\right)\right\}$$

$$U_2=\left\{\left(x^4,x^5\right)\right\}$$

$$V_{1,2}=\left\{\left(x^1,x^4\right),\left(x^1,x^5\right),\left(x^2,x^4\right),\left(x^2,x^5\right),\left(x^3,x^4\right),\left(x^3,x^5\right)\right\}$$

SIR kernel

$$\mathbf{x}^{p} = \mathbf{y}^{(p,0)} \qquad \mathbf{y}^{(p,1)}$$



W_1

Update by the following two equations,

$$\nabla W_1 \leftarrow \eta_1 \frac{\partial E^{att}\left(x^p, x^q\right)}{\partial W_1} - \eta_2 \frac{\partial E^{rep}\left(x^r, x^s\right)}{\partial W_1}$$

$$W_1 \leftarrow W_1 - \nabla W_1$$

In this work, we set,

$$\eta_1 = 0.01, \eta_2 = 0.1$$

This means that the force of repelling is stronger than attracting.

Two-Class Problem

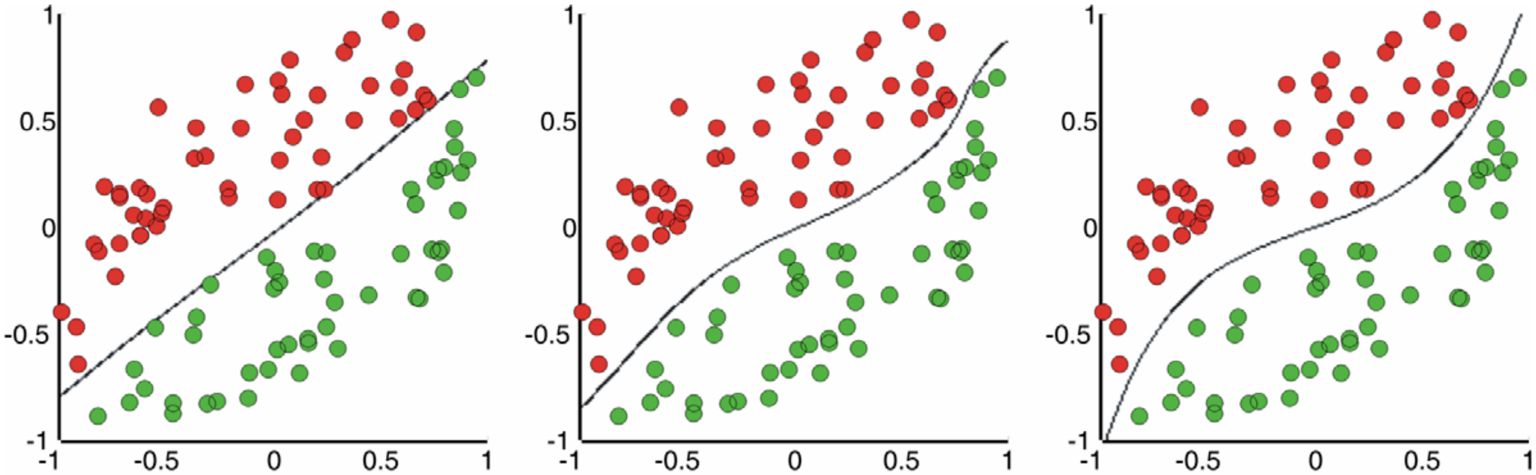
- The border of the data is

$$(x_1)^3 + \frac{1}{10}x_1 = x_2$$

- All input values are in the range [-1,1].

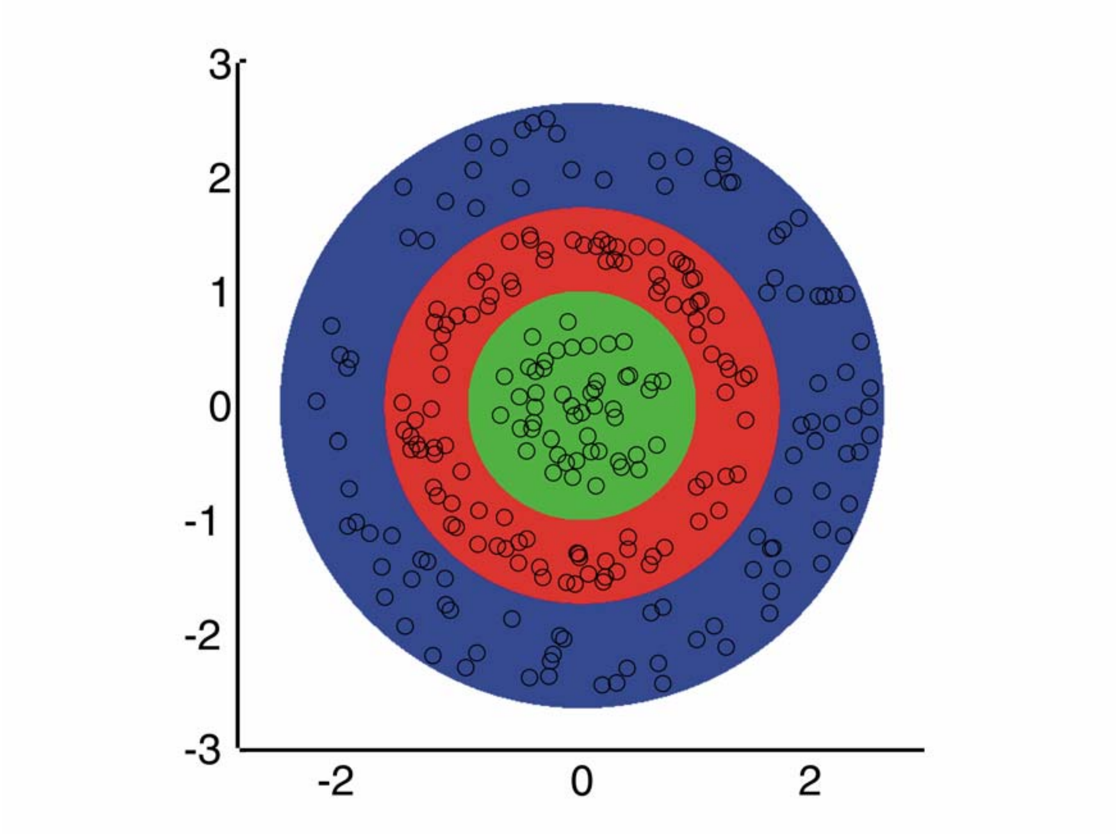
Two-Class Problem

- The number of neuron of SIR kernel is five, n_m=5.
- The supervised BP uses two hidden layers which consists of five neurons, n^{MLP}_1=n^{MLP}_2=5.
- SVM kernel: $K(\mathbf{u},\mathbf{v}) = (\mathbf{u}^T\mathbf{v}+1)^3$



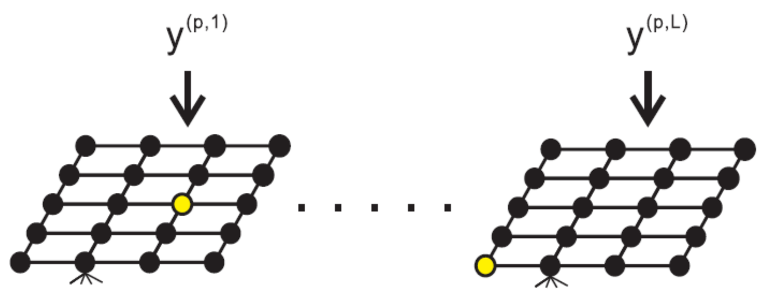
(a) supervised BP (b) SIR kernel (c) SVM

Three-Class Problem

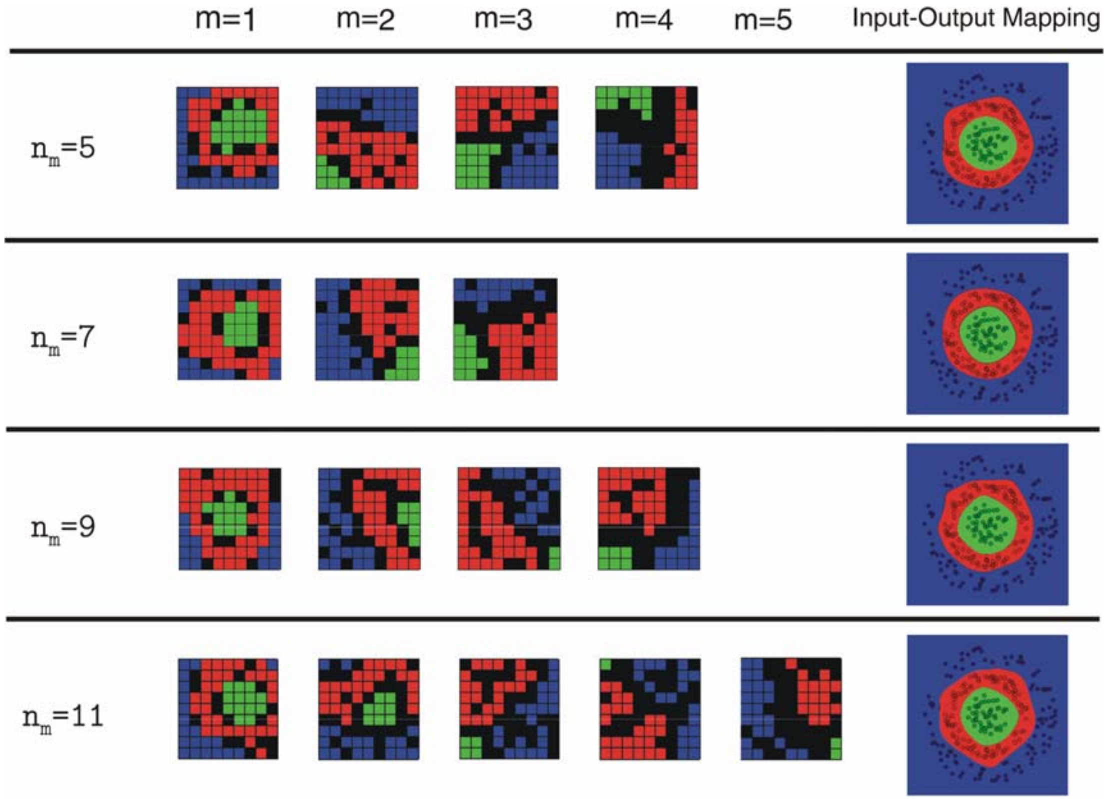


Three-Class Problem

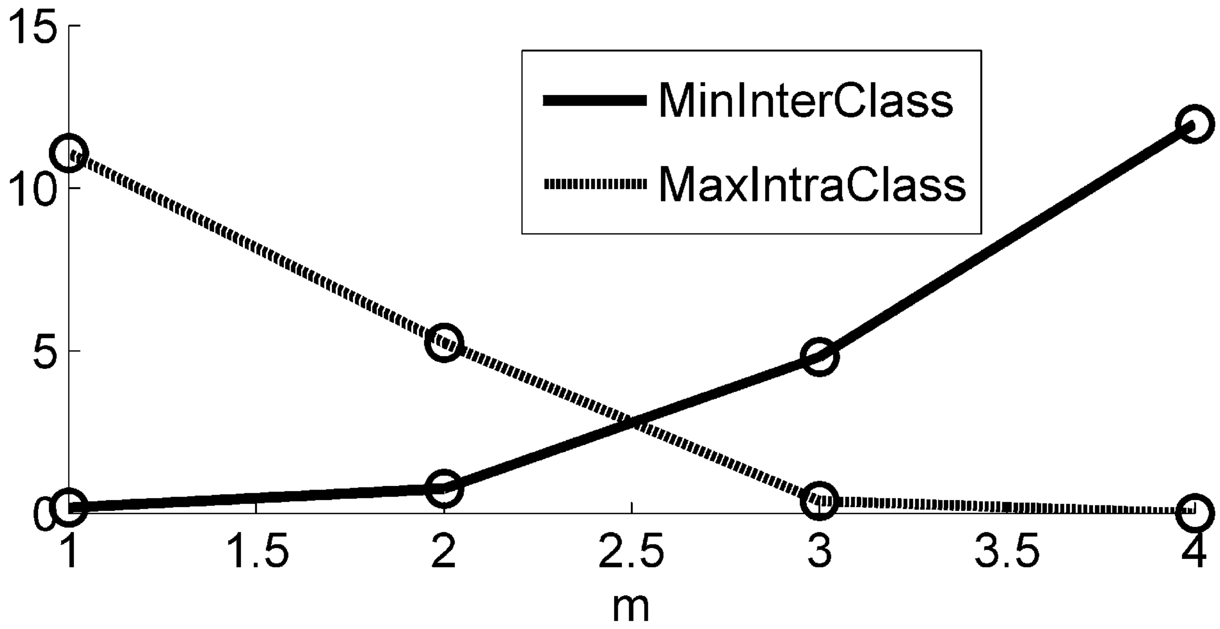
- SOM is used for analyzing the output of each layer. The y is the output of each layer.
- The class color of input pattern are plotted on the winner neuron.



Three-Class Problem



Three-class problem, n_m=5



Real World Data

- Patterns in a whole dataset are divided into 5 partitions.
- The testing accuracy is the average of the 5-fold cross validation.
- The SVM uses a Gaussian kernel.
- The parameters, C and gamma are in the list.

Real World Data

	SIR kernel		SVM		Supervised BP	
	$(n_m, L_{\max})$	(n_1^c, n_2^c)	C	γ	n_1^{MLP}	n_2^{MLP}
iris	(11,5)	(5,3)	50	0.05	20	10
Wisconsin Breast Cancer	(30,7)	(5,1)	50	0.05	30	10
Parkinsons	(20,5)	(5,1)	50	0.05	30	10

Real World Data

	Training Accuracy			Testing Accuracy		
	SIR kernel	Supervised BP	SVM	SIR kernel	Supervised BP	SVM
iris	100%	99.67%	97.50%	97.33%	94.66%	96.00%
Wisconsin Breast Cancer	100%	98.89%	97.53%	96.00%	95.57%	96.42%
Parkinsons	100%	98.33%	99.87%	91.28%	88.20%	92.82%

Summary

- Class to point, guaranteed, $\|Y^L\| = \|C\|$
- Widely separated class points
- Weights by design or training
- Class labels are not used.
- SIR kernel can be used in SVM.
- Hairy network techniques can be used in the calibration sector.
- Suitable for multiple classes problem.

Thank You

Implementation of the MLP Kernel

Cheng-Yuan Liou* and Wei-Chen Cheng
Department of Computer Science and Information Engineering
National Taiwan University
Republic of China
*cyliou@csie.ntu.edu.tw