Multi Class Classification without Multi Class Label

ICLR 2019

Amir Veyseh

Spring 2019

Reading Group, 8 April

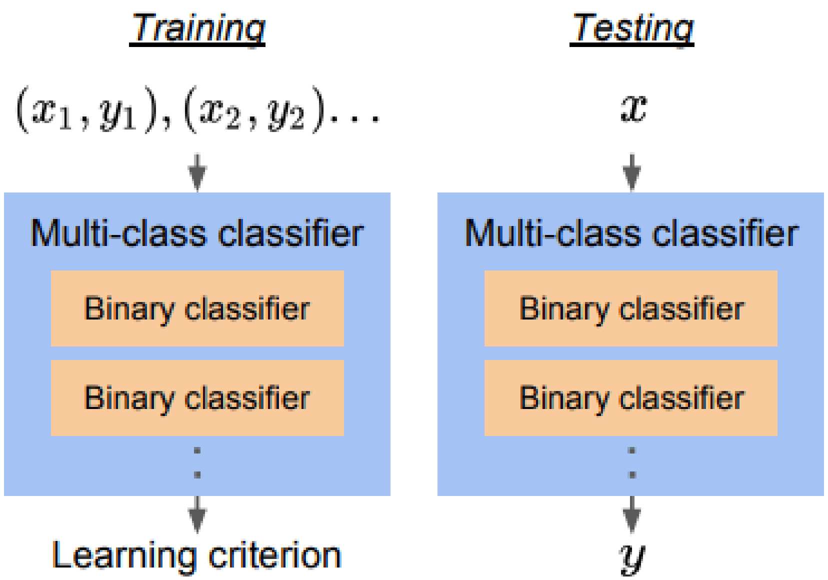
Motivation

- Per class label are expensive to collect
- Labeling require prior knowledge
- Different methods for different types of labeling:
  - Supervised learning
  - Cross-task unsupervised learning
  - Semi-supervised learning

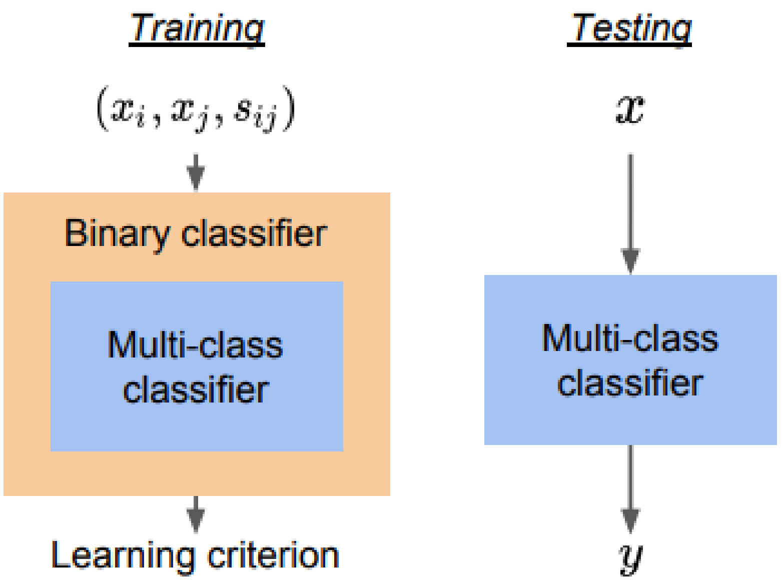
Motivation

- Similarity label instead of class label
  - Reduce problem to binary classification
    - one-vs-one
    - one-vs-all
- Contributions:
  - New Loss function to encapsulate multi-class classifier in binary classifier
  - Experiments on supervised, unsupervised and semi-supervised settings

General Idea



(a) Binary classifier ensemble



(b) Meta classification learning

MCL

$$\mathcal{L}(\theta; \mathbf{X}, \mathbf{Y}, \mathbf{S}) = \mathrm{P}(\mathbf{X}, \mathbf{Y}, \mathbf{S}; \theta) = \mathrm{P}(\mathbf{S}|\mathbf{Y})\mathrm{P}(\mathbf{Y}|\mathbf{X}; \theta)\mathrm{P}(\mathbf{X})$$

$$\mathrm{P}(S_{ij} = 1 | Y_i, Y_j) = 1 \qquad Y_i = Y_j$$

$$\mathrm{P}(S_{ij} = 0 | Y_i, Y_j) = 1 \qquad Y_i \neq Y_j$$

$$\mathrm{P}(\mathbf{S}|\mathbf{Y}) = \prod_{i,j} \mathrm{P}(S_{ij}|Y_i, Y_j)$$

MCL

$$
\begin{aligned}
\mathcal{L}(\theta; \mathbf{X}, \mathbf{S}) &\approx \sum_{\mathbf{Y}} \mathrm{P}(\mathbf{S}|\mathbf{Y})\mathrm{P}(\mathbf{Y}|\mathbf{X};\theta) \\
&\approx \prod_{i,j} \Big(\sum_{Y_i = Y_j} \mathbb{1}[s_{ij} = 1]\mathrm{P}(Y_i|x_i;\theta)\mathrm{P}(Y_j|x_j;\theta) + \\
&\qquad \sum_{Y_i \neq Y_j} \mathbb{1}[s_{ij} = 0]\mathrm{P}(Y_i|x_i;\theta)\mathrm{P}(Y_j|x_j;\theta) \Big).
\end{aligned}
$$

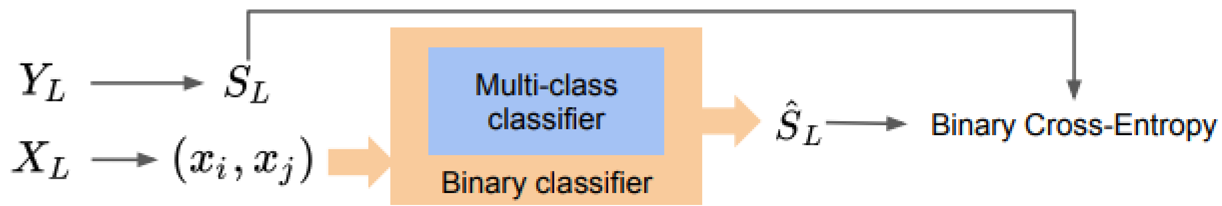
$$
\begin{aligned}
L_{meta}(\theta) &= -\sum_{i,j} \log \Big(\sum_{Y_i = Y_j} \mathbb{1}[s_{ij} = 1]\mathrm{P}(Y_i|x_i;\theta)\mathrm{P}(Y_j|x_j;\theta) + \\
&\qquad \sum_{Y_i \neq Y_j} \mathbb{1}[s_{ij} = 0]\mathrm{P}(Y_i|x_i;\theta)\mathrm{P}(Y_j|x_j;\theta) \Big) \\
&= -\sum_{i,j} s_{ij} \log(f(x_i;\theta)^T f(x_j;\theta)) + (1 - s_{ij}) \log(1 - f(x_i;\theta)^T f(x_j;\theta)).
\end{aligned}
$$

MCL

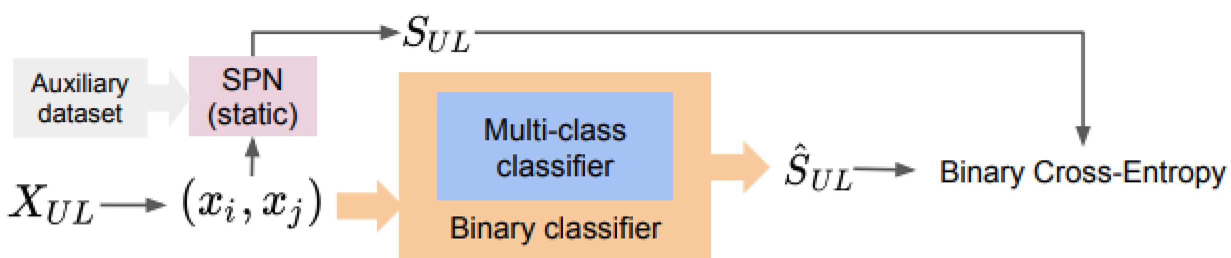
$$g(x_i, x_j, f(\cdot, \theta)) = f(x_i; \theta)^T f(x_j; \theta) = \hat{s}_{ij}$$

$$L_{meta} = -\sum_{i,j} s_{ij} \log \hat{s}_{ij} + (1 - s_{ij}) \log(1 - \hat{s}_{ij})$$

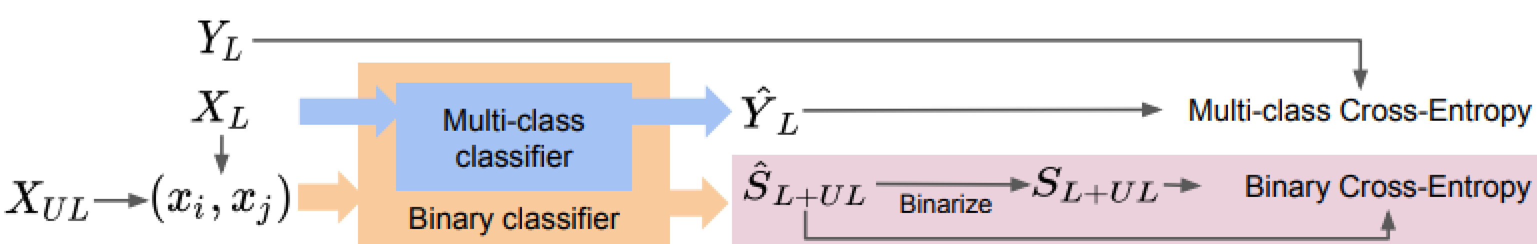
Learning Settings



(a) Supervised learning



(b) Unsupervised transfer learning



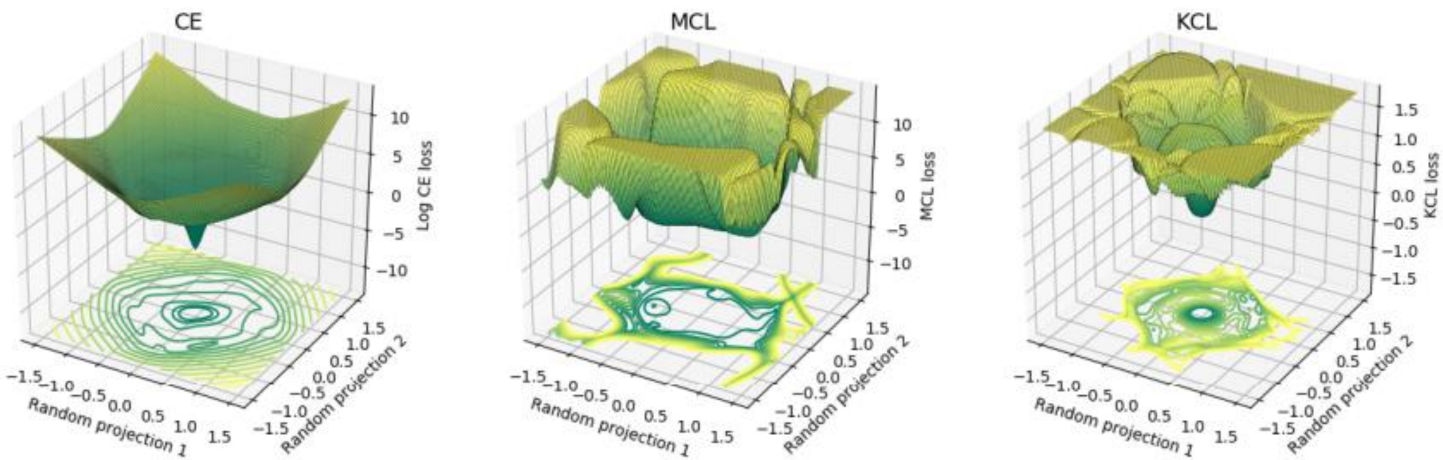
(c) Pseudo-MCL for semi-supervised learning

Experiments

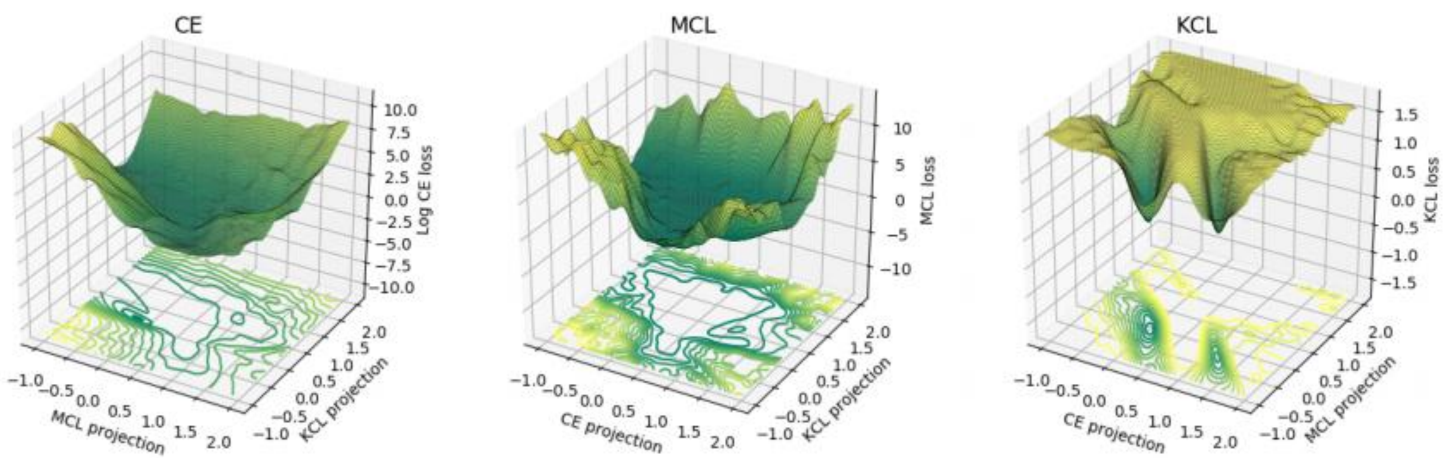
- Supervised Classification
  - Image classification

Dataset	#class	Network	(Class label) CE	(Pairwise label) KCL	MCL
MNIST	10	LeNet	0.6%	**0.5**%	0.6%
CIFAR10	10	LeNet	14.9%	16.4%	15.1%
		VGG8	10.2%	10.2%	10.2%
		VGG11	8.9%	72.2(10.4)%	9.4%
		VGG16	7.6%	*81.1(10.3)%	8.3%
		ResNet18	6.7%	73.8%	6.6%
		ResNet34	6.6%	79.3%	6.3%
		ResNet50	6.6%	79.6%	5.9%
		ResNet101	6.5%	79.9%	**5.6**%
CIFAR100	100	VGG8	**35.4**%	*45.3(40.2)%	36.1%

Experiments



(a) Random projection



(b) Mutual projection

Experiments

- Unsupervised Transfer Learning:
  - Omniglot:
    - 20 images per 1623 handwritten characters
    - Train on 964 character and test on the others
  - ImageNet:
    - Images in 1000 class
    - Train on 882 class and test on the others
- Compare with clustering algorithms
- Two settings:
  - Known number of clusters
  - Unknown number of clusters

Experiments

Method	ACC	ACC(K=100)	NMI	NMI(K=100)
K-means	71.9%	34.5%	0.713	0.671
LSC	73.3%	33.5%	0.733	0.655
LPNMF	43.0%	21.8%	0.526	0.500
KCL	73.8%	65.2%	0.750	0.715
MCL	**74.4%**	**71.5%**	**0.762**	**0.765**

Method	ACC	ACC (K=100)	NMI	NMI (K=100)
K-means (MacQueen et al., 1967)	21.7%	18.9%	0.353	0.464
LPNMF (Cai et al., 2009)	22.2%	16.3%	0.372	0.498
LSC (Chen & Cai, 2011)	23.6%	18.0%	0.376	0.500
ITML (Davis et al., 2007)	56.7%	47.2%	0.674	0.727
SKKm (Anand et al., 2014)	62.4%	46.9%	0.770	0.781
SKLR (Amid et al., 2016)	66.9%	46.8%	0.791	0.760
CSP (Wang et al., 2014)	62.5%	65.4%	0.812	0.812
MPCK-means (Bilenko et al., 2004)	81.9%	53.9%	0.871	0.816
KCL (Hsu et al., 2018)	82.4%	78.1%	0.889	0.874
MCL (ours)	**83.3%**	**80.2%**	**0.897**	**0.893**

Experiments

- Semi-Supervised Setting:
  - CIFAR 10
    - 4k labeled
    - 46k unlabeled

Method	CIFAR10 4k labels
Supervised	$25.4 \pm 1.0\%$
Pseudo-Label	$19.8 \pm 0.7\%$
Π-model	$19.6 \pm 0.4\%$
VAT	$18.2 \pm 0.4\%$
SPN-MCL	$22.8 \pm 0.5\%$
Pseudo-MCL	$\mathbf{18.0} \pm 0.4\%$

Question?