Cluster-dependent Feature Selection by Multiple Kernel Self-organizing Map

Kuan-Chieh Huang, Yen-Yu Lin, and Jie-Zhi Cheng

Research Center for Information Technology Innovation, Academia Sinica, Taiwan

1. Summary

- We propose a new clustering framework called multiple kernel self-organizing map (MK-SOM) that integrates multiple kernel learning (MKL) into the learning procedure of SOM and carries out cluster-dependent feature selection simultaneously.
- **Our approach**
  - Data are characterized by distinct subsets of features or descriptors by considering the generalization of multiple kernel learning for SOM.
  - For cluster-dependent feature selection, the similarity measure of each cluster is represented by a particular combination of the base kernels.
  - To deal with the complex optimization problem, an alternating procedure to optimize both sample coefficient and base kernel coefficient is adopted.

2. The Proposed Approach

2.1 Formulation

- Given $D = \{x_i\}_{i=1...N}$, our goal is to partition D into C clusters.
- The objective function of the SOM can be expressed by

$$E_{SOM} = \sum_{i=1}^{N} \min_j \|x_i - w_j\|^2$$

where sample x_i belongs to the j th cluster, and w_j is the weight vector of the j th neuron on SOM.

- For cluster-dependent feature selection on SOM with MKL:
  - Let φ: $X \rightarrow F$ denote the feature mapping induced by an ensemble kernel, where the w_j lies in the span of data via φ and be weighted by the sample coefficients α.
  - We are to find an optimal convex combination of the base kernels β_m to form the ensemble kernel k.
  - The objective function of MK-SOM can be shown as below

$$\begin{aligned} E_{MK\text{-}SOM} &= \sum_{i=1}^{N} \min_j \Big\|\varphi(x_i) - \sum_{n=1}^{N} \alpha_{j,n}\, \varphi(x_n)\Big\|^2 \\ &= \sum_{i=1}^{N} \min_j \Big[\sum_{m=1}^{M} \beta_m k_m(x_i, x_i) \\ &\qquad - 2 \sum_{n=1}^{N} \alpha_{j,n} \sum_{m=1}^{M} \beta_m k_m(x_n, x_i) \\ &\qquad + \sum_{n=1}^{N} \sum_{n'=1}^{N} \alpha_{j,n}\alpha_{j,n'} \sum_{m=1}^{M} \beta_m k_m(x_n, x_{n'}) \Big] \end{aligned}$$

$$\text{subject to } \sum_{m=1}^{M} \beta_m = 1,\ \beta_m \geq 0\ \forall m$$

- Note that an ensemble kernel is learned for each cluster j.

2.2 Optimization

- An alternating procedure is adopted to optimize both sample coefficient α and base kernel coefficient β iteratively.
- By fixing β, the steepest gradient method is adopted to seek the best α.
  - The partial derivative of the objective function with respect to α can be obtained as

$$\frac{\partial E}{\partial \alpha_{j,n}} = -2\Big[\sum_{m=1}^{M} \beta_m k_m(x_n, x_i) - \sum_{n'=1}^{N} \alpha_{j,n'} \sum_{m=1}^{M} \beta_m k_m(x_n, x_{n'})\Big]$$

  - Hence, the sample coefficient α can be updated by

$$\alpha_{j,n}^{t+1} = \alpha_{j,n}^{t} + \Delta\alpha_{j,n}$$

$$\Delta\alpha_{j,n} = -\eta \cdot \frac{\partial E}{\partial \alpha_{j,n}}$$

- By fixing α, the seeking of the best β is an optimization problem with one additional linear constraint using the reduced gradient descent method.
  - The partial derivative of the objective function with respect to β_m is expressed as

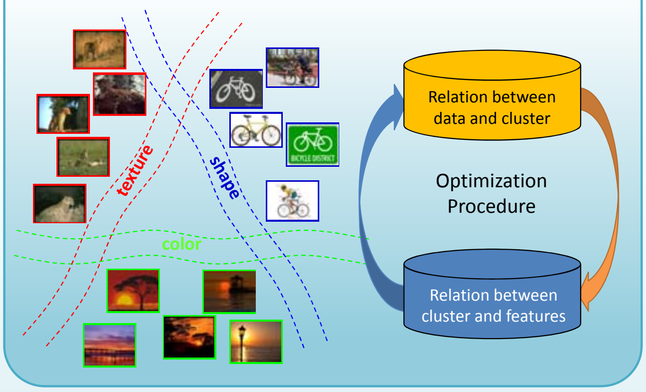
$$\begin{aligned} \frac{\partial E}{\partial \beta_m} &= k_m(x_i, x_i) - 2 \sum_{i=1}^{N} \alpha_{j,n} k_m(x_n, x_i) \\ &\quad + \sum_{n=1}^{N} \sum_{n'=1}^{N} \alpha_{j,n}\alpha_{j,n'}\, k_m(x_n, x_{n'}) \end{aligned}$$

  - The descent direction is obtained by

$$d_m = \begin{cases} 0 & \text{if } \beta_m = 0 \text{ and } \frac{\partial E}{\partial \beta_m} - \frac{\partial E}{\partial d_\mu} > 0 \\ -\frac{\partial E}{\partial \beta_m} + \frac{\partial E}{\partial d_\mu} & \text{if } \beta_m > 0 \text{ and } m \neq \mu \\ \sum_{v \neq \mu, d_v > 0} \big(\frac{\partial E}{\partial \beta_v} - \frac{\partial E}{\partial d_\mu}\big) & \text{for } m = \mu \end{cases}$$

  where μ is selected as the index of the largest component of vector β, for better numerical stability.

Cluster-dependent Feature Selection



3. MK-SOM Training Procedure

- **Input:** Dataset $D = \{x_i\}_{i=1...N}$ in the form of multiple kernels $\{k_m\}_{m=1...M}$
- **Output:** Sample coefficient vectors α_j ;
  Base kernel coefficient vector β ;
- Initial values for α_j and β ;
  - α_j is generated by uniform distribution $[-1, 1]$;
  - β is set as $1/M$ for satisfying constraints;
- **for** $t \leftarrow 1, 2, \ldots, T$ **do**
  - Update α_j by the steepest gradient method;
  - Update β by the reduced gradient method;
    1. Calculate gradient value $\partial E / \partial \beta_m$ and descent direction d_m ;
    2. Iteratively update d_m until convergence as $E(\beta^+) \geq E(\beta)$, where $\beta^+ = \beta + \tau_{max} d$ and τ_{max} is the maximum step size;
    3. Line search along d for appropriate step size τ. $\beta \leftarrow \beta + \tau d$;

end for

return α_j *and* β;

4. Experimental Results

- Two benchmark datasets together with two different schemes of kernel construction are used to evaluate the performance of MK-SOM.
- Clustering performances are evaluated by accuracy (ACC) and normalized mutual information (NMI).

4.1 The Iris Dataset

- The iris dataset consists of 3 classes, each of which contains 50 examples. Data are normalized with their norm in advance.
- The base kernels $k_m(i, j) = \exp(-\|x_i - x_j\|/\sigma_m^2)$, where the number of based kernels is set 5, and σ_m is set as $\{0.2,\ 0.4,\ 0.6,\ 0.8,\ 1.0\}$ respectively.

	K-means	SOM	kSOM	Ours
ACC	0.856	0.887	0.944	**0.977**
NMI	0.742	0.755	0.864	**0.923**

4.2 The Caltech-101 Dataset

- Following the setting in [Dueck et al., ICCV'07], we select the same twenty object categories form the Caltech-101 dataset.
- We randomly pick 30 images from each category to form a set of 600 images.
- Five kinds of image descriptors are implemented, and they result in the following five kernel matrices:
  - GB: Based on the geometric blur descriptor.
  - SIFT: Based on the SIFT descriptor.
  - SS: Based on the self-similarity descriptor.
  - C2: Based on the biologically inspired features.
  - PHOG: Based on the PHOG descriptor.

	K-means + CE	kSOM	Ours
ACC	0.738	0.751	**0.815**
NMI	0.737	0.742	**0.799**

CE: Cluster Ensemble [Strehl et al., JMLR'02]