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, Taipei, Taiwan
kchuang1015@gmail.com, yylin@citi.sinica.edu.tw, jzcheng@ntu.edu.tw

Abstract

Motivated by the fact that data of each cluster are often well captured by distinct features, we propose a clustering approach called multiple kernel self-organizing map (MK-SOM) that integrates multiple kernel learning into the learning procedure of SOM, and carries out cluster-dependent feature selection simultaneously. MK-SOM is developed to reveal the intrinsic relation between features and clusters, and is derived with an efficient optimization procedure. The proposed approach is evaluated on two benchmark datasets, UCI and Caltech-101. The promising experimental results demonstrate its effectiveness.

1. Introduction

Clustering is a versatile technique and has been widely applied to various data analysis problems [1], such as image segmentation, information retrieval, and bioinformatics. Clustering analysis aims to partition the data into a set of coherent groups (clusters). Thus, similarity measures are required to define the coherence. Alas, a universally applicable similarity measure is hardly available. Different similarity measures lead to distinct clustering results. It follows that choosing pertinent similarity measures has been investigated in many clustering methods, e.g., [2].

In addition to similarity measures, attributing data with features (views) is also critical in clustering. As the data labels are often high-level semantic concepts, data of each cluster may be captured by a specific combination of features (views). An example is given in Figure 1, where images with jaguars distinguish themselves from the images of the other classes via the *texture* features. On the other hand, images of class bicycle can be identified with *shapes*, while images of class sunset can be predominantly spotted with *colors*. In this case, a specific cluster/class of data may be superiorly described by particular subsets of features, rather than the whole feature set. In most clustering approaches, features are often treated equally, and hence the underlying relationships between data and the specific subsets of features may be obscured. Unlike most of the prior approaches, a new clustering framework is introduced in this paper to enable data to be characterized by distinct subsets of features or descriptors. It can be achieved by uncovering the underlying relationships between features and data via a *cluster-dependent feature selection* mechanism.

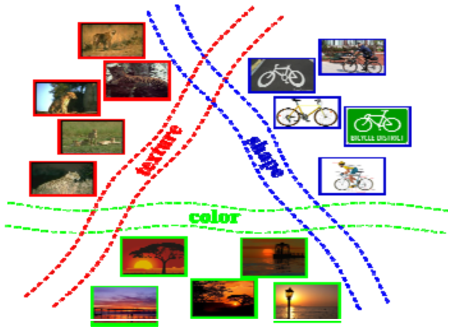

Figure 1. Images from three different categories: sunset, bicycle and jaguar.

In the literature, the most related clustering scheme to our work is multi-view clustering. The optimal solution of multi-view clustering is usually sought by EM algorithm or spectral clustering. For example, Bickel and Scheffer [3] seek the optimal multi-view clustering with a co-EM algorithm. Zhou and Burges [4] extend spectral clustering for multi-view data by generalizing the single-view normalized cut. Mirzaei [5] proposes multi-view agglomerative clustering that extends the hierarchical clustering method to handle multi-view represented objects. Tzortzis and Likas [6] exploit the diverse weightings w.r.t. different views on multi-view convex mixture models. Lin et al. [18] propose an approach to realizing cluster-dependent feature selection. However, these methods suffer from the high computational complexity.

Different from the multiple-view clustering scheme, we investigate the intrinsic relation between features and clusters with the cluster-dependent feature selection mechanism. It is realized with the proposed technique: *multiple kernel self-organizing map* (MK-SOM). SOM [7], inspired by the spatial organization of the brain's functions, is the unsupervised artificial neural networks. SOM is capable of describing high dimensional data with low dimensional manifold [7] and is typically optimized by gradient descent. Thus, its computational complexity is much lower than EM-based or spectral clustering methods. Although several studies explore SOM with kernel methods for nonlinear data clustering [8, 9], it opens issues of how to select an appropriate kernel and its parameters. These kernel choosing and parameter setting issues can be further resolved by *multiple kernel learning* (MKL) [10, 11, 12, 13], in which a kernel machine is learned with multiple base kernels as input.

In this work, we cluster data represented in form of multiple kernels, each of which is generated by either a particular parameter or a specific data descriptor. The proposed MK-SOM integrates MKL into the training procedure of SOM, and carries out cluster-dependent feature selection. That is, the similarity measure of each cluster is derived and represented by one particular combination of these kernels.

2. The MK-SOM framework

In this section, we introduce the proposed approach MK-SOM and its optimization procedure.

2.1. Formulation

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

$$E_{SOM} = \sum_{i}^{N} \min_{j} \left\| x_i - w_j \right\|^2 , \tag{1}$$

where sample x_i currently belongs to the jth cluster, w_j is the weight vector of the jth neuron in SOM and it is also regarded as the clustering center. Our goal is to minimize (1) w.r.t. $\{w_j\}_{j=1}^{C}$ for partitioning data into clusters. Typically a variant of gradient-based methods, called *steepest gradient method*, is used for optimizing the parameters of SOM. Specifically, the weight vector is updated as follows

$$\Delta w_j^{t+1} = \eta^{t+1} \cdot (x_i - w_j) \cdot k(r_j)_j^{t+1} , \tag{2}$$

where t indexes the iteration, and η is the learning rate. Function k in (2) is the *neighborhood* function whose definition is given in the following

$$k(r_j)_j = e^{-(r_j / R^t)^2} . \tag{3}$$

r_j is the distance between the *winner* neuron to neuron j in the output layer. The learning rate is updated by $\eta^{t+1} = \kappa \cdot \eta^t$, and the internal parameter of the neighborhood function k is updated by $R^{t+1} = \lambda \cdot R^t$, where κ and λ are positive constants.

To extend SOM to handle cluster-depend feature selection, we consider its generalization of multiple kernel learning. Let Φ: $X \to F$ denote the feature mapping induced by an *ensemble kernel*. The data are mapped to a high dimensional Hilbert space. In this case, the objective function of MK-SOM becomes

$$E_{MK-SOM} = \sum_{1}^{N} \min_{j} \left\| \phi(x_i) - w_j \right\|^2 , \tag{4}$$

where w_j would lie in the span of data via Φ, i.e.,

$$w_j = \sum_{n=1}^{N} \alpha_{j,n} \phi(x_n) . \tag{5}$$

In Eq. (5), $\{\alpha_{j,n}\}_{j=1}^{N}$ are sample coefficients. It follows that equation (5) can be further expanded as

$$E_{MK-SOM} = \sum_{i=1}^{N} \min_{j} \left\| \phi(x_i) - \sum_{n=1}^{N} \alpha_{j,n} \phi(x_n) \right\|^2 \tag{6}$$

$$= \sum_{i=1}^{N} \min_{j} [\phi^T(x_i)\phi(x_i) - 2\sum_{n=1}^{N} \alpha_{j,n} \phi^T(x_i)\phi(x_n) + \sum_{n=1}^{N}\sum_{n'=1}^{N} \alpha_{j,n}\alpha_{j,n'} \phi^T(x_n)\phi(x_{n'})] \tag{7}$$

$$= \sum_{i=1}^{N} \min_{j} [k(x_i, x_i) - 2\sum_{n=1}^{N} \alpha_{j,n} k(x_i, x_n) + \sum_{n=1}^{N}\sum_{n'=1}^{N} \alpha_{j,n}\alpha_{j,n'} k(x_n, x_{n'})]. \tag{8}$$

Eq. (8) is obtained from (7) with the kernel trick.

Like the seminar work of MKL, e.g., [10, 11, 12, 13], our MKL formulation is to find an optimal way to linearly combine the given *base kernels*. Namely, the ensemble kernel k in (8) is a *convex combination* of the base kernels, i.e.,

$$k(x_i, x_j) = \sum_{m=1}^{M} \beta_m k_m(x_i, x_j) \quad \text{subject to} \sum_{m=1}^{M} \beta_m = 1,\ \beta_m \geq 0\ \forall \text{m} , \tag{9}$$

where β_m is a base kernel coefficient. M is the numbers of base kernels. The resulting objective function is

$$E_{MK-SOM} = \sum_{i=1}^{N} \min_{j} [\sum_{m=1}^{M} \beta_m k_m(x_i, x_i) - 2\sum_{n=1}^{N} \alpha_{j,n} \sum_{m=1}^{M} \beta_m k_m(x_n, x_i) + \sum_{n=1}^{N}\sum_{n'=1}^{N} \alpha_{j,n}\alpha_{j,n'} \sum_{m=1}^{M} \beta_m k_m(x_n, x_{n'})] \quad \text{subject to} \sum_{m=1}^{M} \beta_m = 1,\ \beta_m \geq 0\ \forall \text{m} . \tag{10}$$

Note that an ensemble kernel is learned for each cluster j. It indicates that cluster-dependent feature (kernel) selection will be carries out.

2.2. Optimization

Optimization problem in (10) is too complex to be solved directly. We hence adopt an alternating procedure to optimize both the sample coefficient *α* and base kernel coefficient *β* iteratively. Specifically, one of *α* and *β* is first optimized while the other is fixed, and their roles are switched. The procedure is repeated until convergence. Figure 2 gives the pseudo-codes of the training procedure for our MK-SOM.

Optimizing *α*. By fixing *β*, the *steepest gradient* method is adopted to seek the best *α* at current iteration *t* with

$$\alpha_{j,n}^{t+1} = \alpha_{j,n}^{t} + \Delta\alpha_{j,n} \tag{11}$$

$$\Delta\alpha_{j,n} = -\eta \cdot \frac{\partial E}{\partial \alpha_{j,n}} \ . \tag{12}$$

Due to the symmetric properties of kernels, i.e., $k(x_i, x_j) = k(x_j, x_i)$, the partial derivative of the objective function w.r.t. *α* can be obtained as

$$\begin{aligned} &\frac{\partial E_{MK-SOM}}{\partial \alpha_{j,n}} \\ &= -2 \cdot [\sum_{m=1}^{M} \beta_m k_m(x_n, x_i) - \sum_{n'=1}^{N} \alpha_{j,n'} \sum_{m=1}^{M} \beta_m k_m(x_n, x_{n'})] \ . \end{aligned} \tag{13}$$

It follows that the sample coefficient *α* can be updated by

$$\begin{aligned} &\alpha_{j,n}^{t+1} = \alpha_{j,n}^{t} \\ &+ 2\eta \cdot [\sum_{m=1}^{M} \beta_m k_m(x_n, x_i) - \sum_{n'=1}^{N} \alpha_{j,n'} \sum_{m=1}^{M} \beta_m k_m(x_n, x_{n'})] \ . \end{aligned} \tag{14}$$

Optimizing *β*. By fixing *α*, the seeking of the best *β* is an optimization problem with one additional linear constraint. Hence, we employ the *reduced gradient descent* method [14], which is shown to be able to deal with additional constraints in the procedure of gradient descent effectively. The partial derivative of the object function w.r.t. *β* is expressed as

$$\begin{aligned} \frac{\partial E_{MK-SOM}}{\partial \beta_m} &= k_m(x_i, x_i) - 2\sum_{n=1}^{N} \alpha_{j,n} k_m(x_n, x_i) \\ &+ \sum_{n=1}^{N} \sum_{n'=1}^{N} \alpha_{j,n} \alpha_{j,n'} k_m(x_n, x_{n'}) \ . \end{aligned} \tag{15}$$

Then, the optimal *β* at iteration *t* can be sought with the procedures described in Figure 2. Please refer to [14] for the details of the reduced gradient descent.

3. Experimental results

Two benchmark datasets together with two different schemes of kernel construction are used to evaluate the performance of MK-SOM. The first dataset is the *iris* data from UCI [15], where the base kernels are built with different hyper-parameters in the RBF function. The second one is the Caltech-101 data set [16], in which five different image descriptors are adopted, and hence five corresponding kernels are then constructed.

In all experiments, we set the number of clusters to the number of classes in the ground truth. Two criteria, *accuracy* (ACC) and *normalized mutual information* (NMI), are used for evaluating clustering performance.

```
Input:  Dataset $D = \{x_i\}_{i=1}^{N}$ in the form of multiple
        kernels $\{k_m\}_{m=1}^{M}$;
Output: Sample coefficient vectors $\alpha_j$;
        Base kernel coefficient vector $\beta$;
Initial values for $\alpha_j$ and $\beta$;
  $\alpha_j$ is generated by uniform distribution [-1, 1];
  $\beta$ is set as 1/M for satisfying constraints;
for t ← 1, 2, …, T do
  1. Update $\alpha_j$ by the steepest gradient method in Eq. (14);
  2. Update $\beta$ by the reduced gradient method;
    2.1. Find Index $I$ with largest component of $\beta$;
    2.2. Let $\beta=(\beta_B, \beta_N)$; Vector $a$ is constraint
         coefficients. $\beta_{B=}\{a_k:k\in I\}$, $\beta_{N=}\{a_k:k\notin I\}$;
    2.3. Calculate gradient value $\nabla \beta_B, \nabla \beta_N$ by Eq. (15);
    2.4. Calculate reduced gradient $r$ by
         $r(\beta_N) = \nabla\beta_N - (\beta_B^{-1}\beta_N)\cdot\nabla\beta_B$;
    2.5. Calculate possible gradient $d$ by
         $d_N = \{ \begin{matrix} -\beta_N r, & \text{if } r > 0 \\ -r, & \text{if } r \le 0 \end{matrix}$;
         $d_B = -\beta^{-1} N d_N$
    2.6. Line search along $d$ for appropriate step
         size τ. $\beta \leftarrow \beta + \tau\beta$;
end for
return $\alpha_j$ and $\beta$;
```

Figure 2. The training procedure of MK-SOM.

3.1. The Iris dataset

The iris dataset consists of three classes, each of which contains fifty examples. Data are normalized with their norm in advance. The base kernels are then established with different parameters, i.e.,

$$k_m(i, j) = \exp(-\|x_i - x_j\| / \sigma_m^2) \tag{16}$$

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.

We compare our approach with three clustering algorithms, including *k*-means, SOM, and kernel SOM (kSOM). Note that kSOM works with each of the five kernels. The best performance of the five kernels is reported. It can be observed that our approach can take the information embedded in the five kernels into account and leads to a significant improvement.

Table 1. The performances, in forms of ACC and NMI, of four different clustering methods in the iris dataset.

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

3.2. The Caltech-101 dataset

The Caltech-101 dataset is adopted in the second set of experiments. The large intra-class variations make clustering over these data very challenging. We follow the setting of [19]. The same 20 categories from the Caltech-101 dataset are selected, and we randomly pick 30 images from each category to form a set of 600 images. Five different image descriptors are adopted for establishing base kernels, including four shape descriptors (geometric blur, SIFT, self-similarity, and PHOG) and one biologically inspired feature (FH)

Two baselines are adopted, i.e., kernel *k*-means and kernel SOM. Each baseline works with one kernel at a time. The five clustering results of each baseline are merged by *cluster ensembles* [17], one of the state-of-the-art approaches to clustering result combination. Unlike cluster ensembles that fuse multiple clustering results in a global fashion, our approach achieves cluster-dependent feature selection over the multiple descriptors to recover the underlying structure of each cluster. As shown in Table 2, the proposed approach outperforms the two baselines.

Table 2. The performances, in forms of ACC and NMI, of three clustering methods in the Caltech-101 dataset.

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

4. Conclusion

In this work, we propose MK-SOM that integrates multiple kernel learning into SOM, and carries out cluster-dependent feature selection. The alternating optimization procedure is adopted for deriving both the sample and kernel coefficients in an efficient way. The experimental results on two benchmark datasets demonstrate that our approach effectively uncovers the underlying relationships between features and clusters.

Acknowledgments. The work is supported in part by grants NSC 100-2218-E-001-004.

References

[1] R. Xu and D. Wunsch II. Survey of clustering algorithms, *IEEE Trans. Neural Networks*, 2005.

[2] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering, *Neural Information Processing System*, 2004.

[3] S. Bickel and T. Scheffer. Multi-view clustering, i*n Proc. of Int. Conf. on Data Mining*, 2004.

[4] D. Zhou and C. Burges. Spectral clustering and transductive learning with multiple views, *Int. Conf. on Machine Learning*, 2007.

[5] H. Mirzaei. A novel multi-view agglomerative clustering algorithm based on ensemble of partitions on different views, *Int. Conf. on Pattern Recognition*, 2010.

[6] G. F. Tzortzis and A. C. Likas. Multuple view clustering using a weighted combination of exemplar-based mixture models, *IEEE Trans. Neural Networks*, 2010.

[7] T. Kohonen. Self-organized formation of topologically correct feature maps, *Biological Cybernetics*, 1982.

[8] B. Romain, J. Bertrand, R. Fabrice, and V. Nathalie. Batch kernel SOM and related Laplacian methods for social network analysis, *Neurocomputing*, 2008.

[9] K. Lau, H. Yin, and S. Hubbard. Kernel self-organising maps for classification, *Neurocomputing*, 2006.

[10] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the smo algorithm, *Int. Conf. on Machine Learning*, 2004.

[11] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. More efficiency in multiple kernel learning, *Int. Conf. on Machine Learning*, 2007.

[12] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. SimpleMKL, *Journal of Machine Learning Research*, 2008.

[13] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh. Multiple kernel learning for dimensionality reduction, *IEEE Trans. Pattern Analaysis and Machine Intelligence*, 2011.

[14] P. Wolfe. Methods of nonlinear programming, *Recent Advances in Mathematical Programming*, R.L. Graves and P. Wolfe (Eds.), 1963.

[15] A. Asuncion and D. J. Newman. UCI machine learning repository, *Irvine, CA: University of California, School of Information and Computer Science*, 2007.

[16] L. Fei-Fei, R. Fergus, P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories, *Int. Conf. on Computer Vision and Pattern Recognition*, 2004.

[17] A. Strehl, J. Ghosh. Cluster ensembles – A knowledge reuse framework for combing multiple partitions, *Journal of Machine Learning Research*, 2002.

[18] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh. Clustering complex data with group-dependent feature selection, *European Conf. on Computer Vision*, 2010.

[19] D. Dueck and B. Frey. Non-metric affinity propagation for unsupervised image categorization, *Int. Conf. on Computer Vision*, 2007.