Kernelized Evolutionary Distance Metric Learning for Semi-Supervised Clustering

Wasin Kalintha,[1] Satoshi Ono,[2] Masayuki Numao,[3] Ken-ichi Fukui[3]
[1]Graduate School of Information Science and Technology, Osaka University, 1-5 Yamadaoka Suita Osaka 565-0871 Japan, wasin@ai.sanken.osaka-u.ac.jp, +81-6-6879-8426
[2]Graduate School of Science and Engineering, Kagoshima University, 1-21-40 Kohrimoto, Kagoshima-city 890-0065, Japan, ono@ibe.kagoshima-u.ac.jp
[3]The Institute of Scientific and Industrial Research, Osaka University, 8-1 Mihogaoka Ibaraki, Osaka 567-0047 Japan, {surname}@ai.sanken.osaka-u.ac.jp

Introduction

Many research studies on distance metric learning (DML) reiterate that the definition of distance between two data points substantially affects clustering tasks. Recently, variety of DML methods have been proposed to improve the accuracy of clustering by learning a distance metric (Moutafis, Leng, and Kakadiaris 2016); however, most of them only perform a linear transformation, which yields insignificant to non-linear separable data. This study proposes a DML method which provides an integration of kernelization technique with Mahalanobis-based DML. Thus, non-linear transformation of the distance metric can be performed. Moreover, a cluster validity index is optimized by an evolutionary algorithm. The empirical results on semi-supervised clustering suggest the promising result on both synthetic and real-world data set.

Related Work

Kernelized Kmeans Clustering

Kernelized Kmeans clustering (K-KMN) (Dhillon, Guan, and Kulis 2004) is an enhancement of Kmeans clustering (KMN), that can extract clusters that are non-linearly separable in the original data space by applying a proper nonlinear mapping function (kernel) to a higher dimensional feature space.

Given a data set $\mathcal{D} = \{\mathbf{x}_i = (x_{i,1}, \cdots, x_{i,v}) \in R^v\}_{i=1}^N$, let the k^{th} cluster $C_k \in \mathbf{C}$. Using non-linear function $\phi(\mathbf{x})$, the objective function of K-KMN is defined as:

$$\min \sum_{C_k \in \mathbf{C}} \sum_{\mathbf{x}_i \in C_k} \parallel \boldsymbol{\pi}_k - \phi(\mathbf{x}_i) \parallel_2^2 \tag{1}$$

The centroid of cluster C_k, $\boldsymbol{\pi}_k$ is defined as:

$$\boldsymbol{\pi}_k = \frac{\sum_{\mathbf{x}_i \in C_k} \phi(\mathbf{x}_i)}{\mid C_k \mid} \tag{2}$$

Since the mapping function $\phi(\mathbf{x})$ is hard to obtain, thus kernel function $K(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{y})$ is calculated instead.

Copyright © 2017, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

$$\parallel \boldsymbol{\pi}_k - \phi(\mathbf{x}_i) \parallel_2^2 = \frac{\sum_{\mathbf{x}_j, \mathbf{x}_l \in C_k} K(\mathbf{x}_j, \mathbf{x}_l)}{\mid C_k \mid^2} - \frac{2 \sum_{\mathbf{x}_j \in C_k} K(\mathbf{x}_i, \mathbf{x}_j)}{\mid C_k \mid} + K(\mathbf{x}_i, \mathbf{x}_i) \tag{3}$$

Evolutionary Distance Metric Learning (EDML)

EDML, proposed by Fukui et al. (Fukui et al. 2013), is a DML technique based on the Mahalanobis-based distance. Distance metric matrix $\mathbf{M}$ is a variable to be learned in order to maximize a cluster validity index as in Eq. 4 which is optimized by self-adapting control parameters and generalized opposition- based differential evolution (GOjDE), a state-of-the-art differential evolution technique. Note that $\mathbf{M}$ must be a symmetric positive semi-definite matrix (PSD) to satisfy the distance propositions. Let $m_{i,j} \in \mathbf{M}$ and $d_{i,j}^2$ is a distance metric for the corresponding semi-supervised clustering.

$$\begin{aligned} &\text{Maximize } Eval(Clustering(d_{i,j}^2)), \\ &\text{s.t. } |m_{k,k}| \geq \sum_{l(k \neq l)} |m_{k,l}|, \\ &0 < m_{k,k} \leq 1,\ -1 \leq m_{k,l} \leq 1\ (k \neq l). \end{aligned} \tag{4}$$

Proposed Method

Unlike other kernelized DMLs (Moutafis, Leng, and Kakadiaris 2016) which formulate a penalty function for constraints, i.e., must-link and cannot-link, into an objective function, the proposed kernelized evolutionary distance metric learning (K-EDML) can directly improve cluster validity index as an objective function. This method can perform non-linear transformation while in EDML cannot. Further, the symmetric PSD matrix $\mathbf{M}$ can be decomposed into $\mathbf{M} = \mathbf{L}^t\mathbf{L}$ by Cholesky decomposition, where $\mathbf{L}$ denotes an upper triangular matrix. Mahalanobis distance in EDML can be rewritten as:

$$\begin{aligned} d_{i,j}^2 &= \parallel \mathbf{x}_i - \mathbf{x}_j \parallel_{\mathbf{M}}^2 \\ &= (\mathbf{x}_i - \mathbf{x}_j)^t \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^t \mathbf{L}^t \mathbf{L} (\mathbf{x}_i - \mathbf{x}_j) \\ &= (\mathbf{L}\mathbf{x}_i - \mathbf{L}\mathbf{x}_j)^t (\mathbf{L}\mathbf{x}_i - \mathbf{L}\mathbf{x}_j) = \parallel \mathbf{L}\mathbf{x}_i - \mathbf{L}\mathbf{x}_j \parallel_2^2 \end{aligned} \tag{5}$$

We substitute $\mathbf{x}_i$ with $\mathbf{L}\mathbf{x}_i$ into the K-KMN objective function Eq. 1 and used as $Clustering()$ in Eq. 4.

$$\min \sum_{C_k \in \mathbf{C}} \sum_{x_i \in C_k} \| \boldsymbol{\pi}_k - \phi(\mathbf{L}\mathbf{x}_i) \|_2^2 \qquad (6)$$

Experimental Evaluation

We study the performance of $\mathbf{M}$ from K-EDML by comparing it to other clustering algorithms, i.e., KMN, K-KMN, and EDML. Each method is performed for 5000 trials with random initialization, and evaluated using external criteria, i.e., class information (label). Cluster validity index is used as $Eval()$ in Eq. 4. Unlike conventional cluster validity that can only evaluate individual cluster quality, weighted pairwise F-measure ($wPFM$) (Fukui et al. 2013) is applied here to evaluate the overall cluster structure and neighbor cluster relations. We selected non-linearly separable data sets, i.e., artificial data set (Rings) and Pima Indians Diabetes data set (Pima) obtained from UCI Machine Learning Repository[1] in order to present the effectiveness of the proposed K-EDML. Note that appropriate kernel and hyperparameters selection is out of the scope of this work and that is the reason we try only the degree-2 polynomial kernel with $c = 0$. Thus, $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}\mathbf{y})^2$. Label sampling rate for EDML and K-EDML is set to 30% of the data.

Artificial Data Set

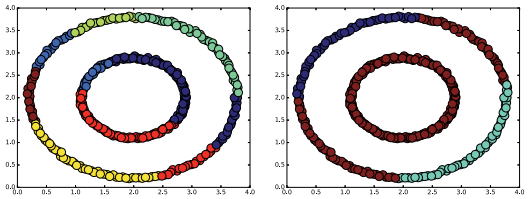
The experiment is conducted on the synthetic data to confirm the explicit performance of the proposed method. It consists of two circles as shown in Figure 1. Each circle represents one class, each class contains 200 instances with normal distribution on the specific radius. We set the cluster size to 10 and neighbor size to 5. The cluster size has to be larger than the actual class size in order to evaluate using the $wPFM$. Table 1 shows the evaluation result in average and standard deviation of each clustering algorithm. The K-EDML outperforms all other clustering algorithms with 99% confidence level via paired t-test. Also, visualization of the clustering result is presented in Fig. 1. Obviously, KMN and EDML cannot perform well due to the non-linearly separable data sets. Thus, EDML cannot improve the cluster validity index score or performs even worse when data is non-linearly separable. Although KKMN with hyperparameter tuning can easily cluster this data set, it fails to archive fine clustering when the hyperparamter is inappropriate. On the other hand, K-EDML takes advantage of class label utilization, thus a more flexible kernel data space in K-EDML is obtained.

Table 1: Clustering results on each algorithm ($wPFM$)

	KMN	KKMN	EDML	K-EDML
Rings	0.2824 ± 0.2529	0.2972 ± 0.2661	0.2749 ± 0.2462	$\mathbf{0.3739 \pm 0.3351}$
Pima	0.2703 ± 0.2422	0.2815 ± 0.2520	0.3090 ± 0.2772	$\mathbf{0.4107 \pm 0.3677}$

[1]https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes

Complementary material:
http://www.ai.sanken.osaka-u.ac.jp/files/AAAI17-supplyment.pdf



(a) EDML (b) K-EDML

Figure 1: Visualization of cluster assignment in input space. (KMN and KKMN is omitted due to the space limitation)

Pima Indians Diabetes Data Set

K-EDML is then applied on a real-world data set, Pima, contains 2 classes with 8 dimensions and 768 instances. The dimension is normalized to $\mu = 0$ and $\sigma = 1$. We set cluster and neighbor size to 20 and 5 respectively. The evaluation results in average and standard deviation of each clustering algorithm are shown in Table 1. K-EDML still archived similar results like in the previous data set. According to the paired t-test, K-EDML outperforms all other methods with 95% confidence. The improvement of $wPFM$ in Table 1 clearly illustrated the performance of K-EDML. Furthermore, EDML outperforms KMN and KKMN in this dataset, which confirms the advantage of utilizing class information in clustering on both EDML and K-EDML.

Conclusion

This paper proposes a non-linear transformation of distance metric learning. Experimental results on synthetic and real-world data set, Pima, empirically demonstrates the drawback of EDML in non-linearly separable input space and illustrate the benefit of kernel function to the proposed K-EDML method due to its superior results to other clustering algorithms for semi-supervised clustering. Lastly, we aim to proceed this research by applying other kernel function with hyperparameter tuning during optimization and also improving the computational efficiency of the proposed method.

Acknowledgement

This work is partially supported by the cooperative research program of Network Joint Research Center for Materials and Devices.

References

Dhillon, I. S.; Guan, Y.; and Kulis, B. 2004. Kernel k-means: Spectral clustering and normalized cuts. In *Proceedings of KDD '04*, 551–556. ACM.

Fukui, K.; Ono, S.; Megano, T.; and Numao, M. 2013. Evolutionary distance metric learning approach to semi-supervised clustering with neighbor relations. In *Proc. of ICTAI '13*, 398–403.

Moutafis, P.; Leng, M.; and Kakadiaris, I. A. 2016. An overview and empirical comparison of distance metric learning methods. *IEEE Transactions on Cybernetics* (99):1–14.