Automatic Kernel Parameter Tuning of KSC for Video Category Classification

Taiyo Mineo and Haruhisa Takahashi

Graduate School of the University of Electro-Communications
1-5-1 Chofugaoka, Chofu, Tokyo 182-8585, Japan
Phone/FAX: +81-42-443-5291
E-mail: takahasi@ice.uec.ac.jp

Abstract

Here we focus on website video-tag category clustering using kernel spectral clustering (KSC) [1], which has not been reported so far. KSC is a nonlinear clustering method used for the spectral clustering of data in a kernel feature space. Unfortunately KSC is intrinsically affected by the selection of kernel parameters, and this hinders complete automatic learning. This study proposes a unified learning method which enables automatic KSC optimal kernel parameter tuning based on a graph Laplacian representation of the clustering. This method is well suited to applications that require speedy and fully automatic analysis, since it can automatically find the optimal parameter, and generate the optimally connected components of the graph Laplacian. We demonstrate the application of our method to the above problem and show its excellent performance through computer experiments.

1. Introduction

Clustering is one of the most important unsupervised learning methods due to its broad application range. Although methods such as the k-means method and Ward’s method are some of the popular, their classification accuracy is insufficient for real-world problems. An important problem is to increase the classification ability while maintaining or improving the usability.

As clustering targets are not always partially linearly classifiable, nonlinear clustering based on kernel methods has been proposed. KSC [1] can sensitively detect similar features of clusters due to a skilled combination of spectral clustering with kernel principal component analysis (kernel PCA). Unfortunately, it is necessary to determine suitable or optimal kernel parameters in KSC since it is based on the radial basis function (RBF) kernel. The RBF with too small kernel parameters is hypersensitive to noise, whereas the RBF with too large parameters overlooks cluster features [2]. In [1], parameter tuning is realized by implementing an exhaustive search to find the best clustering result. This method is not appropriate for real-world applications that require speedy processing, yet optimization of the parameters is critical for such applications since the evaluation process is based on an external criterion.

This study proposes a unified learning method which enables automatic KSC optimal kernel parameter tuning based on a graph Laplacian representation of the clustering. The proposed kernel parameter-tuning method directly evaluates spectral clustering eigenvalues, and thus can find a suitable separation of the undirected graph of the graph Laplacian. The computational cost of parameter tuning is very small despite the method being based on a simple steepest descent method in light of the nonexistence of local optimal solutions.

We apply our fully automatic KSC approach to movie classification based on movie tags. Although previous works have focused on the description of tag network figures, there have been no works on classification analysis by high-precision clustering. It should be worthwhile analyzing the continually changing data stream of a movie site community using our method, which should be useful for analyzing big data.

In computer experiments we analyze movie tag data in various periods, then examine the changes in the data stream of the movie community. We show that our method achieves better purity than the k-means method in real-time, although it is slower than the k-means method.

2. Methods

We first overview KSC in this section, then the automatic parameter-tuning method is described.

2.1 KSC

The KSC primal problem is to construct projections $e^{(l)}$ ($l = 1, \cdots, n_c$) which have the maximum variance in a feature space by applying an LS-SVM and weighted-kernel PCA simultaneously. This problem is formulated as

$$\min_{w^{(l)}, e^{(l)}, b_l} \frac{1}{2} \sum_{l=1}^{n_c} w^{(l)T} w^{(l)} - \frac{1}{2N} \sum_{l=1}^{n_c} \gamma_l e^{(l)T} V e^{(l)}$$

subject to : $e^{(l)} = \Phi w^{(l)} + b_l 1_N$

where $w^{(l)}$ is the coefficients vector, $\gamma_l \in \mathbb{R}^+$ are regularization constants, $\Phi$ is the training feature matrix, $V$ is a user-defined weight matrix, and $1_N$ is the $N$-dimensional vector.
of ones. From the Lagrangian $L$ of Eq (1), the Karush-Kuhn-Tucker (KKT) conditions are given by

$$\begin{align*}
\frac{\partial L}{\partial \alpha} &= 0 \rightarrow w^{(l)} = \Phi^T \alpha^{(l)} \cdots (I) \\
\frac{\partial L}{\partial b} &= 0 \rightarrow \alpha^{(l)} = \frac{1}{N} \sum_l V e^{(l)} \cdots (II) \\
\frac{\partial L}{\partial \theta} &= 0 \rightarrow \alpha^{(l)T} 1_N = 0 \cdots (III) \\
\frac{\partial L}{\partial \gamma} &= 0 \rightarrow e^{(l)} = \Phi w^{(l)} + b 1_N \cdots (IV)
\end{align*}$$

Eliminating $w^{(l)}$ and $e^{(l)}$ and expressing the KKT conditions in terms of the dual variables $\alpha^{(l)}$ lead to the dual eigenvalue problem with the eigenpair \{\(\lambda_l = \frac{\sqrt{\gamma_l}}{\gamma_l}, \alpha^{(l)}\)\}:

$$V \left( I - \frac{1}{N} 1_N 1_N^T V \right) \Omega \alpha^{(l)} = \lambda_l \alpha^{(l)}$$

(3)

where \(\Omega = \Phi \Phi^T\) is interpreted as the Gram (kernel) matrix. Let \(V^{-1}\) be the degree matrix (the sum of the edge weights incident at nodes) and \(\Omega\) be the similarity matrix. We construct the normalized graph Laplacian as \(\Omega\), which is interpreted as a random-walk transition matrix. If the graph contains \(k > 1\) connected components, there are \(k\) eigenvalues of \(\Omega \Omega\) close to 1. Thus Eq (3) can be transformed by \(V \Omega \alpha^{(l)} = \alpha^{(l)}\):

$$\alpha^{(l)} - \frac{1}{\sqrt{\gamma_l}} 1_N V 1_N \alpha^{(l)} = \lambda_l \alpha^{(l)}$$

$$\Rightarrow \alpha^{(l)} = \lambda_l \alpha^{(l)} \quad (\because \text{KKT conditions Eq (2-III)})$$

Therefore, Eq (3) has the n,\(\gamma\)-dimensional eigenspace, which is spanned by eigenvectors of the maximum eigenvalue (\(= 1\)). We set \(n_k = k - 1\) since \((k - 1)\)-dimensional space \(k\) clusters are linearly separable. The eigenvectors and the training data projections have the same sign pattern due to the KKT conditions Eq (2-II) since \(\gamma_l\) and the diagonal of \(V\) are positive. By focusing on the sign pattern of the projection (encoding vector), we can perform clustering. A codebook can be formed from the \(k\) most commonly occurring encoding vectors in the training data. When we evaluate out-of-sample data, its encoding vector is compared with the codewords in the codebook and it is assigned to the cluster with minimal Hamming distance.

KSC performs spectral clustering in a dual eigenspace using the RBF kernel with kernel parameter \(\theta\). We use the expression \(K(x_i, x_j)\) as the kernel function, where

$$K(x_i, x_j) = \exp \left[ -\left( \frac{d(x_i, x_j)}{\theta} \right)^2 \right]$$

and where \(d(\cdot, \cdot)\) is the distance function between the input data \(x_i\) and \(x_j\).

### 2.2 Optimal criterion for eigenvalues

We apply the kernel parameter learning method [2] to the symmetric normalized graph Laplacian \(L_{\text{sym}} = I - V \frac{1}{2} \Omega V \frac{1}{2}\) (I : identity matrix). Let \(L_{rw} = I - V \Omega\). Then \(L_{sym}\) is related to \(L_{rw}\) as described in the following theorem [3]:

**Theorem 1** (Number of connected components and spectra of \(L_{\text{sym}}\) and \(L_{rw}\)). Let \(G\) be an undirected graph with non-negative weights. Then the multiplicity \(k\) of the eigenvalue 0 of both \(L_{rw}\) and \(L_{sym}\) equals the number of connected components \(A_1, \cdots, A_k\) in \(G\).

From Theorem 1, if we construct \(L_{sym}\) with the optimal connected components, then \(V \Omega\) can be constructed under an optimal condition. As the eigenvalues \(\mu_l\) of \(L_{sym}\) are not explicitly given, the partial derivatives of \(\mu_l\) with respect to the kernel parameter \(\theta\) are obtained using the implicit function theorem.

Let us define \(F(\mu, \theta) = \mu I - L_{sym}(\theta)\). Then the characteristic equation of \(L_{sym}\) can be described as \(\det \{F(\mu, \theta)\} = 0\).

By the implicit function theorem, if \(\frac{\partial}{\partial \mu} \det \{F(\mu, \theta)\} \neq 0\), then the implicit function \(f_l(\theta) = \mu_l\) exists and is differentiable. Then the partial derivative of \(f_l\) is described by

$$\frac{\partial f_l(\theta)}{\partial \theta} = \frac{\partial \mu_l}{\partial \mu} \frac{\partial \mu_l}{\partial \theta} = \frac{\partial \det \{F(\mu, \theta)\}}{\partial \mu_l} \frac{\partial \det \{F(\mu, \theta)\}}{\partial \mu_l} = \frac{\partial \det \{F(\mu, \theta)\}}{\partial \mu_l} \frac{\partial L_{sym}}{\partial \theta} \frac{\partial L_{sym}}{\partial \mu_l}$$

$$\text{where } \tilde{F} \text{ is the cofactor matrix of } F. \text{ In order to simplify the calculation of the cofactor matrix in Eq (5) we apply the following theorem [2].}

**Theorem 2** (Cofactor matrix of \(\mu_l I - L_{sym}\)). Let \(\mu_1, \cdots, \mu_N\) be the eigenvalues of \(L_{sym}\) and \(v_1, \cdots, v_N\) be the correspond eigenvectors. Then the cofactor matrix of \(F = \mu_l I - L_{sym}\) can be calculated by the following equation.

$$\tilde{F} = \prod_{i=1}^N \left( \mu_i - \mu_l \right) v_i v_i^T$$

(6)

Then, Eq (5) can be rewritten as

$$\frac{\partial \mu_l}{\partial \theta} = \frac{\partial \mu_l}{\partial \mu} \frac{\partial \mu_l}{\partial \theta} = \frac{\partial \det \{F(\mu, \theta)\}}{\partial \mu_l} \frac{\partial \det \{F(\mu, \theta)\}}{\partial \mu_l} = \frac{\partial \det \{F(\mu, \theta)\}}{\partial \mu_l} \frac{\partial L_{sym}}{\partial \theta} v_i v_i^T$$

(7)

We can use Eq (7) as the partial derivatives of the eigenvalue with respect to the kernel parameter.
2.3 Optimal parameter

In order to obtain the optimal parameter, we need to define a criterion function \( E(\theta) \), which is constructed by evaluating some aspect of the connected components for the current parameter.

In Fig 1 the eigenvalues of \( L_{sym} \) are shown in ascending order. Note that all the eigenvalues are in the range \([0, 1]\) and that \( \mu_1 = 0 \). In this spectrum the first \( k \) eigenvalues are close to 0 (see Theorem 1). In KSC, we should find \( k \) connected components in the undirected graph. Therefore, optimal eigenvalues should be an eigengap between \( \mu_k \) and \( \mu_{k+1} \) close to 1 (\( |\mu_k - \mu_{k+1}| \to 1 \)). From Fig 1, we can see that the maximum eigengap gives optimal condition.

We slightly modify this condition to define a smooth criterion function: the sum of the first \( k \) eigenvalues should be an eigengap between \( \mu_k \) and \( \mu_{k+1} \) close to 1 (\( |\mu_k - \mu_{k+1}| \to 1 \)). From Fig 1, we can see that the maximum eigengap gives optimal condition.

We slightly modify this condition to define a smooth criterion function: the sum of the first \( 2(k - 1) \) eigenvalues is equal to nearly \( k - 1 \) under the optimal condition. According to this criterion, \( E(\theta) \) and its derivative are defined as follows:

\[
E(\theta) = \frac{1}{k-1} \left[ \sum_{i=2}^{k} \mu_i^2 + \sum_{i=k+1}^{2k-1} (1 - \mu_i)^2 \right] \tag{8}
\]

\[
\frac{\partial E(\theta)}{\partial \theta} = \frac{2}{k-1} \left[ \sum_{i=2}^{k} \mu_i \frac{\partial \mu_i}{\partial \theta} + \sum_{i=k+1}^{2k-1} (\mu_i - 1) \frac{\partial \mu_i}{\partial \theta} \right] \tag{9}
\]

The criterion function \( E(\theta) \) is normalized in \((0, 1)\); thus, the minimum value of \( E(\theta) \) is interpreted as the optimal status.

We apply the steepest descent method to minimize \( E(\theta) \). The update rule at step \( t \) for parameter \( \theta_t \) is written as

\[
\theta_{t+1} = \theta_t - \eta \frac{\partial E(\theta_t)}{\partial \theta_t} \tag{10}
\]

where \( \eta \) is the learning rate. To make the computation process stable, we include the size of the previous update in the update rule:

\[
\theta_{t+1} = \theta_t - \left[ \eta \frac{\partial E(\theta_t)}{\partial \theta_t} + \alpha \frac{\partial E(\theta_{t-1})}{\partial \theta_{t-1}} \right] \tag{11}
\]

where \( \alpha \) is a coefficient smaller than \( \eta \).

3. Experiments

In this section we examine the features of the criterion function and then apply our method to website video-tag category clustering.

3.1 Features of \( E(\theta) \)

We first use the data of the Iris dataset\(^1\) to investigate the features of \( E(\theta) \). We use the euclidean distance function \( d(x_i, x_j) = ||x_i - x_j|| \) for the kernel function. In Fig 2 \( E(\theta) \) with the purity is shown and in Fig 3 the derivative of \( E(\theta) \) is shown for \( k = 3 \), where the purity is defined by

\[
\text{purity} = \frac{1}{N} \sum_{i=1}^{k} \max |C_i \cap A_j| \tag{12}
\]

where \( \{C_1, \ldots, C_k\} \) are the clustered results for \( N \) data points and \( \{A_1, \ldots, A_k\} \) are target clusters. From this example, we

---

\(^1\)Available at http://archive.ics.uci.edu/ml/datasets/Iris
can see that $E(\theta)$ has one optimal solution which gives good purity. For other cases, the behavior of the function is similar. In this example, the optimal parameter that minimizes $E(\theta)$ is $0.868$. For the Iris data the total computational time for parameter learning and the clustering process was $0.655[\text{sec}]$ (Fedora-19 (VMware) on Windows 7, Intel i7 3.50GHz, 16GB, implemented by Ruby), which is longer than the simple k-means method but is sufficiently fast for real applications. We show the clustering result in Fig 4, and the clustering purity was $91.0[\times 10^{-2}]$. We applied the method in [1] to the same problem and obtained an optimal parameter of $0.058$ with a computational time $67.8[\text{sec}]$ as well as a purity of $59.7[\times 10^{-2}]$. From this experiment, the proposed method exhibits far better performance than the conventional method [1].

### 3.2 Movie classification

We collected data from Nico Nico Douga\(^1\), choosing categories from the major ones (Games, Anime, Music, Entertainment, Government, Sports, Science). Each video datum is given by a vector of video-tag indices (0 or 1). The cosine similarity function $d(x_i, x_j) = 1 - \frac{x_i^T x_j}{||x_i|| ||x_j||}$ is used as the distance in the RBF kernel function. We first compared our method with the k-means method. Results are shown in Table 1 for two categories. From this table we can see that KSC attains far better results than the k-means method. In Fig 5 maintains average purities are compared for 2 to 7 categories. While KSC high purity when the number of categories is increased, for the k-means method the purity decreases with increasing number of categories.

\(^{1}\text{http://www.nicovideo.jp/}\)

| KSC | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|-----|---|---|---|---|---|---|---|
| Games (1) | - | 80.7 | 79.2 | 75.9 | 80.0 | 80.4 | 73.9 |
| Anime (2) | 95.8 | - | 79.0 | 75.0 | 81.1 | 78.7 | 79.1 |
| Music (3) | 94.3 | 95.1 | - | 80.2 | 82.5 | 75.9 | 79.7 |
| Entertainment (4) | 92.8 | 94.8 | 99.1 | - | 78.7 | 78.5 | 71.4 |
| Government (5) | 97.9 | 96.7 | 95.2 | 93.1 | - | 76.7 | 78.0 |
| Sports (6) | 84.7 | 96.7 | 91.1 | 92.2 | 90.8 | - | 78.0 |
| Science (7) | 99.9 | 97.3 | 94.0 | 93.7 | 92.1 | 92.4 | - |

Figure 4: Clustering result for Iris data

Figure 5: Average purity obtained by KSC and k-means methods

### 4. Conclusions

We showed through video category classification tasks that KSC with automatic parameter tuning is effective in terms of both computational time and classification capability. A notable property of the proposed method compared with the k-means method is that the classification capability does not deteriorate with increasing number of categories. A future study will be to generalize our method to hierarchical clustering to enable more detailed category classification.

Acknowledgment

This work was supported in part by a Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology, Japan (No. 24500165).

References

[1] C. Alzate and J. A. K. Suykens: Hierarchical kernel spectral clustering, Neural Networks, Vol. 35, pp.21-30, 2012.
[2] T. Nogayama, H. Takahashi and M. Muramatsu: Generalization of kernel PCA and automatic parameter tuning, Proc. ANZIIS, pp.173-178, 2004.
[3] U. von Luxburg: A tutorial on spectral clustering, Technical report, No. TR-149, Max-Planck-Institut fur Biologische Kybernetik, 2007.