Extraction of Dimension Reduced Features from Empirical Kernel Vector

Takio Kurita and Yayoi Harashima

Department of Information Engineering
Hiroshima University
1-4-1 Kagamiyama, Higashi-Hiroshima, 739-8527, Japan

Abstract. This paper proposes a feature extraction method from the given empirical kernel vector. We show the necessary condition for the feature extraction mapping to make the trained classifier by using the linear SVM with the extracted feature vectors equivalent to the one obtained by the standard kernel SVM. The proposed feature extraction mapping is defined by using the eigen values and eigen vectors of the Gram matrix. Since the eigen vector problem of the Gram matrix is closely related with the kernel Principal Component Analysis, we can extract a dimension reduced feature vector. This feature extraction method becomes equivalent to the kernel SVM if the full dimension is used. The proposed feature extraction method was evaluated by the experiments using the standard data sets. The cross-validation values of the proposed method were improved and the recognition rates were comparable with the original kernel SVM. The number of extracted features was very low compared to the number of features of the kernel SVM.

1 Introduction

Support vector machine (SVM) [15,12,3,7] has been successfully applied to many pattern recognition problems such as object detection[5] and image classification[4] etc. The nonlinear classifier with good generalization can be constructed by using kernel-trick and margin maximization.

However, the dimension of the empirical kernel vector in the kernel SVM increases as the number of training samples increases. Especially for big data, this makes the computation of the learning algorithm intractable. Also the generalization ability of the trained classifier probably decreases because the number of parameters increases as the number of training samples increases. It is well known that the complexity of the model used in the learning and the intrinsic dimension to describe the target classification problem should be the same to get the classifier with good generalization.

To reduce the difference between the complexities of the learning model and the the target classification problem, Nishida et al. proposed a method to select the important kernel features by using Boosting [10]. Also Nishida et al. proposed an algorithm called RANSAC-SVM in which the subsets of the training samples were randomly generated and the best subset was selected [11].

C.K. Loo et al. (Eds.): ICONIP 2014, Part II, LNCS 8835, pp. 9–16, 2014.

In this paper, we propose a method to extract new feature vector from the given empirical kernel vector by using kernel Principal Component Analysis (the kernel PCA).

At first, we show the necessary condition which has to be satisfied to make the classifier obtained by using the linear SVM with the extracted feature vector from the empirical kernel vector in the case of full dimension equivalent to the one trained by using the standard kernel SVM. Then a feature extraction mapping from the given empirical kernel vector is derived. The feature extraction mapping can be defined by using the eigen values and eigen vectors of the Gram matrix.

The eigen vector problem of the Gram matrix is closely related with the kernel PCA [1,13]. The eigen vectors corresponding to the first largest eigen values are the principal components and extract the dominant information from the Gram matrix. By combining the feature extraction mapping by Gram matrix and the dimension reduction by the kernel PCA, we can design feature extraction method in which the dominant information of the given empirical kernel vector is extracted but the unnecessary details are neglected. This feature extraction becomes equivalent to the kernel SVM if the full dimension is used.

By the experiments using the standard data sets, we found that the very few features were enough to achieve good recognition rates for test samples by using the proposed dimension reduced feature vectors. These results shows that the empirical kernel vector includes redundant information and there is margin to reduce the number of dimension.

As the related works, general theory of the kernel PCA whitening in kernel-based methods is shown in [13]. A relation between the kernel PCA and the least squares SVM (LS-SVM) is shown in [14]. Q. Chen et al. proposes a combination of the kernel PCA and LS-SVM and applied to time series prediction [2]. In this paper we experimentally evaluate the effect of the dimension reduction by using the kernel PCA whitening for the case of the kernel SVM. Also it is reported that the whitening using the covariance matrix of the HOG features can improve the recognition performance when it is used as the input of the linear SVM in [6]. The tendency of our experimental results agrees with the results of the whitening of the HOG features.

2 Feature Extraction from Empirical Kernels

In this paper we extract dimension reduced feature vector from the given empirical kernel vector by using a linear mapping. Then the extracted new feature vectors are used as the input of the linear SVM. We want to make the obtained classifier without dimension reduction identical to the one obtained by the original kernel SVM. To consider the constraints which should be satisfied by this linear mapping, we will briefly review the linear and kernel SVM.

2.1 Linear and Kernel SVM

The linear SVM determines the separating hyperplane with a maximal margin by using the given training samples $\{< \boldsymbol{x}_i, t_i > | i = 1, \ldots, n\}$, where $\boldsymbol{x}_i$ is the input feature vector and $t_i = \in \{+1, -1\}$ is the class label of the i-th sample. Then the classification function of the linear SVM is given as

$$y = \mathrm{sign}(\boldsymbol{w}^T \boldsymbol{x} - h), \tag{1}$$

where $\boldsymbol{w}$ and h are the weight vector and the threshold, respectively. The function sign(u) is a sign function which outputs 1 when $u > 0$ and outputs -1 when $u \leq 0$. The soft-margin SVM is defined as an optimization problem for the following evaluation function

$$L(\boldsymbol{w}, \boldsymbol{\xi}) = \frac{1}{2}||\boldsymbol{w}||^2 + \gamma \sum_{i=1}^{n} \xi_i, \tag{2}$$

under the constraints $\xi_i \geq 0, \quad t_i(\boldsymbol{w}^t \boldsymbol{x}_i - h) \geq 1 - \xi_i, \quad i = 1, \ldots, n$, where ξ_i is the measure of the error for the training sample $\boldsymbol{x}_i$. The dual problem is obtained as the optimization problem that maximizes the object function

$$L_D(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j t_i t_j \boldsymbol{x}_i^T \boldsymbol{x}_j \tag{3}$$

under the constraints $\sum_{i=1}^{n} \alpha_i t_i = 0, \quad 0 \leq \alpha_i \leq \gamma, \quad i = 1, \ldots, n$.

By solving this optimization problem, the optimal classification function can be expressed as

$$y = \mathrm{sign}(\sum_{i \in S} \alpha_i^* t_i \boldsymbol{x}_i^T \boldsymbol{x} - h^*), \tag{4}$$

where S is a set of support vectors and α_i^* and h^* are the optimal solutions.

By using the kernel-trick, this linear SVM can be extended to nonlinear (kernel SVM). In the kernel SVM, input vectors are mapped to higher dimensional feature space by non-linear function $\boldsymbol{\phi}(\boldsymbol{x})$ and the linear SVM is applied to the mapped features. Since the linear SVM depends only on the inner products of the input vectors, we can define the object function as

$$L_D(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \ \alpha_j t_i t_j K(\boldsymbol{x}_i, \boldsymbol{x}_j), \tag{5}$$

where $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{\phi}(\boldsymbol{x}_i)^T \boldsymbol{\phi}(\boldsymbol{x}_j)$ is the kernel function. Usually the kernel function $K(\boldsymbol{x}, \boldsymbol{y})$ is defined a priori. The polynomial function and the Radial Basis function are often used as the kernel function.

Then the optimal classification function of the kernel SVM can be derived as

$$y = \mathrm{sign}(\sum_{i \in S} \alpha_i^* t_i \boldsymbol{\phi}(\boldsymbol{x}_i)^T \boldsymbol{\phi}(\boldsymbol{x}) - h^*) = \mathrm{sign}(\sum_{i \in S} \alpha_i^* t_i K(\boldsymbol{x}_i, \boldsymbol{x}) - h^*). \tag{6}$$

2.2 Feature Extraction from Empirical Kernel Vector

The classification function of the kernel SVM given in equation (6) determines the separating hyperplane on the n dimensional feature vector

$$\boldsymbol{k}(\boldsymbol{x}) = (K(\boldsymbol{x}_1, \boldsymbol{x}), \ldots, K(\boldsymbol{x}_n, \boldsymbol{x}))^T. \tag{7}$$

This means that this n dimensional feature vector include enough information to construct the classification function of the kernel SVM. We call this n dimensional feature vector the empirical kernel vector.

To extract effective features from this empirical kernel vector, we consider the following linear feature extraction

$$\boldsymbol{g}(\boldsymbol{x}) = U^T \boldsymbol{k}(\boldsymbol{x}). \tag{8}$$

Then we use this new feature vector $\boldsymbol{g}(\boldsymbol{x})$ as the input of the linear SVM.

By substituting the $\boldsymbol{x}_i$ with the new feature $\boldsymbol{g}(\boldsymbol{x}_i)$ in the equation (3), we have

$$L_D(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{j=1}^{n} \alpha_j t_j \mathit{\Gamma} K U U^T \boldsymbol{k}(\boldsymbol{x}_j), \tag{9}$$

where $K = (K(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j=1}^{n}$ and $\mathit{\Gamma} = \mathrm{diag}(\alpha_1 t_1, \ldots, \alpha_n t_n)$. The matrix K is known as the kernel Gram matrix.

Similarly, the optimal classification function becomes

$$y = \mathrm{sign}(\mathit{\Gamma}^* K U U^T \boldsymbol{k}(\boldsymbol{x}) - h^*), \tag{10}$$

where $\mathit{\Gamma}^* = \mathrm{diag}(\alpha_1^* t_1, \ldots, \alpha_n^* t_n)$.

To get the same object function and the classification function with the kernel SVM, the coefficient matrix U must be satisfy the condition $UU^T = K^{-1}$. Since the kernel Gram matrix K is real symmetric, we can compute the U by using the eigen values λ_i and the corresponding eigen vectors $\boldsymbol{a}_i$ of the kernel Gram matrix K. The eigen values and the corresponding eigen vectors of the kernel Gram matrix K are given by

$$KA = A\mathit{\Lambda} \tag{11}$$

where $\mathit{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ is the diagonal matrix with the eigen values and $A = (\boldsymbol{a}_1 \cdots \boldsymbol{a}_n)$ is the matrix of eigen vectors. Thus the coefficient matrix U in the feature extraction mapping to make the obtained classifier coincident with the kernel SVM can be given by

$$U = A\mathit{\Lambda}^{-\frac{1}{2}}, \tag{12}$$

where $\mathit{\Lambda}^{-1/2} = \mathrm{diag}(\frac{1}{\sqrt{\lambda_1}}, \frac{1}{\sqrt{\lambda_2}}, \ldots, \frac{1}{\sqrt{\lambda_n}})$. Since the matrix A is orthogonal, we can confirm the condition of the inverse matrix as

$$K(UU^T) = KA\mathit{\Lambda}^{-\frac{1}{2}}\mathit{\Lambda}^{-\frac{1}{2}}A^T = A\mathit{\Lambda}\mathit{\Lambda}^{-1}A^T = AA^T = I. \tag{13}$$

Then the feature extraction which gives the same results with the kernel SVM is given by

$$\boldsymbol{g}_{SVM}(\boldsymbol{x}) = \mathit{\Lambda}^{-\frac{1}{2}} A^T \boldsymbol{k}(\boldsymbol{x}). \tag{14}$$

This feature extraction is closely related with the kernel PCA and we can extract the dimension reduced features in terms of the kernel principal components.

2.3 Relation with Kernel Principal Component Analysis

Nonlinear extension of PCA using the kernel-trick is known as the kernel PCA. In the kernel PCA, the input vectors are also mapped to the higher dimensional feature space by non-linear function $\boldsymbol{\phi}(\boldsymbol{x})$ and the linear PCA is applied to the mapped features.

For the given data set $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, the kernel PCA computes the principal score as

$$\boldsymbol{y}(\boldsymbol{x}) = U^T \boldsymbol{\phi}(\boldsymbol{x}). \tag{15}$$

Since the coefficient matrix can be represented by the linear combinations of the mapped feature vectors as

$$U = \sum_{j=1}^{n} \boldsymbol{\phi}(\boldsymbol{x}_j)\boldsymbol{\alpha}_j^T, \tag{16}$$

the principal score vector can be given by

$$\boldsymbol{y}(\boldsymbol{x}) = \sum_{j=1}^{n} \boldsymbol{\alpha}_j \boldsymbol{\phi}(\boldsymbol{x}_j)^T \boldsymbol{\phi}(\boldsymbol{x}) = \sum_{j=1}^{n} \boldsymbol{\alpha}_j K(\boldsymbol{x}_j, \boldsymbol{x}). \tag{17}$$

The optimal solution can be obtained by taking the L eigen vectors $\tilde{A} = (\boldsymbol{\alpha}_1 \cdots \boldsymbol{\alpha}_L)$ of the kernel Gram matrix K corresponding to the L largest eigen values $\lambda_1, \ldots, \lambda_L$. The eigen vector equation for kernel PCA is given by

$$K\tilde{A} = \tilde{A}\tilde{\Lambda}, \tag{18}$$

where $\tilde{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_L)$.

By comparing the eigen vector equation (18) for the kernel PCA and the eigen vector equation (11) for the feature extraction from the empirical kernels, it is noticed that they are the same. Since the kernel PCA can extract dominant information from the data set and neglect the unnecessary details by taking the principal components, we can construct new dimension reduced features by taking the L eigen vectors of the kernel Gram matrix K corresponding to the L largest eigen values as

$$\boldsymbol{g}_{PCA}(\boldsymbol{x}) = \tilde{\Lambda}^{-\frac{1}{2}} \tilde{A}^T \boldsymbol{k}(\boldsymbol{x}). \tag{19}$$

This feature vector $\boldsymbol{g}_{PCA}(\boldsymbol{x})$ can extract the dominant information of the training data set and neglect the unnecessary details. Also this feature vector can produce almost same result with the kernel SVM when this feature vector is used as the input of the linear SVM. Especially the result becomes the same as the kernel SVM if the full dimension, namely $L = n$, is used.

3 Experiments

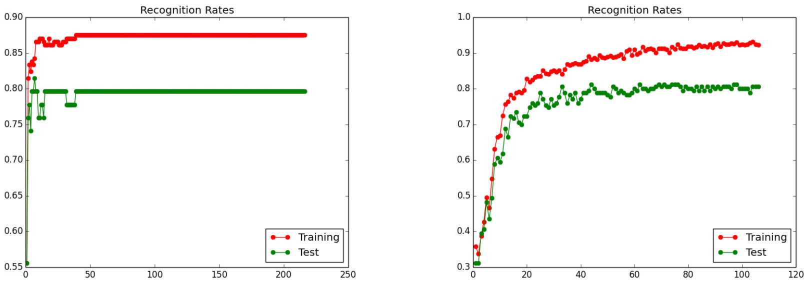
The effectiveness of the proposed feature extraction defined in the equation (19) was evaluated by using seven standard data sets (**heart**, **iris**, **vowel**, **breast-cancer**, **glass**, and **vehicle**) from LIBSVM data sets [8]. The number of classes,

Table 1. The cross validation values (%), recognition rates (%) and the dimension of extracted feature vector for the standard data sets

data set		K-SVM	FE-PCA
heart	CV	85.33 (2.06)	**86.16** (1.32)
	training	**87.41** (2.06)	86.94 (1.32)
	test	78.52 (5.67)	**78.70** (5.03)
	dim.	216	18.6 (20.17)
iris	CV	97.50 (0.88)	**97.58** (1.00)
	training	97.66 (1.23)	**97.83** (1.05)
	test	**95.33** (2.81)	93.67 (4.83)
	dim.	120	4.8 (1.93)
vowel	CV	98.58 (0.45)	**98.70** (0.49)
	training	**99.98** (0.07)	99.95 (0.01)
	test	**98.58** (0.80)	**98.58** (0.67)
	dim.	677	168.7 (67.15)
breast-cancer	CV	97.20 (0.35)	**97.29** (0.42)
	training	**97.33** (0.36)	97.31 (0.37)
	test	**97.23** (1.74)	**97.23** (1.81)
	dim.	546	7.8 (8.66)
glass	CV	72.88 (3.20)	**74.10** (2.76)
	training	**92.75** (5.20)	89.59 (6.39)
	test	67.21 (8.09)	**68.14** (7.44)
	dim.	171	52.6 (29.61)
vehicle	CV	85.30 (1.10)	**85.89** (0.98)
	training	**93.24** (0.80)	91.57 (0.90)
	test	**83.41** (2.33)	83.29 (2.86)
	dim.	170	89.8 (10.76)

the number of samples, and the number of features are (2, 270, 13), (3, 150, 4), (11, 528, 10), (2, 683, 10), (6, 214, 9), and (4, 846, 18) respectively. For classification experiments, each data set was randomly divided into a training set (80% of all samples) and a test set (remaining samples). We performed 10 times with different partitions of training and test samples and the average and the standard deviation were measured. The Radial Basis functions $K(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{||\boldsymbol{x}-\boldsymbol{y}||^2}{2\sigma^2}\right)$ was used as the kernel function.

Table 1 shows the cross-validation values and the recognition rates for training samples and test samples of the proposed feature extraction methods, namely the feature extraction from empirical kernel vectors by using PCA (denoted as FE-PCA). In the proposed feature extraction method, the extracted feature vector was classified by using the linear SVM. The recognition rates of the standard kernel SVM (denoted as K-SVM) are also shown in Table 1 for comparison. In these experiments, the Radial Basis function is used as the kernel function. The parameters of the linear and the kernel SVM, i.e. the soft margin parameter γ and the kernel parameter σ^2 were determined by 10-fold cross validation (For the data



(a) heart

(b) vehicle

Fig. 1. Relation between the recognition rates and the extracted dimension

set **glass**, 5-fold cross validation was used because of the shortage of the samples of each class). Since the recognition rate depends on the number of extracted features in the proposed feature extraction methods, the best dimension was selected by the cross validation. The average and the standard deviation of the selected dimension are also shown in Table 1 (denoted as dim.).

From Table 1, the cross-validation values of the proposed feature extraction method gives a little bit better results for all data sets than the standard SVM while the recognition rates for training and test samples are comparable. This means that the proposed dimension reduction method can keep the generalization performance of the kernel SVM.

The number of extracted features is very low compared to the number of original features of the kernel SVM. For example, 546 was reduced to 7.8 for **breast-cancer**. This means that the empirical kernel vector includes very redundant information and there is margin to reduce the number of dimension.

Figure 1 shows the relation between the recognition rates and the number of dimension of the extracted feature vector for **heart** and **vehicle** data sets. It is noticed that there is almost flat regions from low to high dimension. This also shows that information included in the empirical kernel vector is very redundant and the proposed feature extraction method can extract intrinsic information from the empirical kernel vector.

Since the number of features of the kernel SVM increases as the number of training samples increases, the difference between the number of features of the kernel SVM and the intrinsic dimension of the target classification problem becomes large for large data. It is expected that the proposed feature extraction method can be used to reduce this gap.

Acknowledgment. The preliminary experiments on this work was mainly done by Mr. Kohei Takahashi. This work was partly supported by KAKENHI (23500211).

References

1. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
2. Chen, Q., Chen, X., Wu, Y.: Optimization Algorithm with Kernel PCA to Support Vector Machines for Time Series Prediction. Journal of Computers 5(3), 380–387 (2010)
3. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and other Kernel-based Learning Methods. Cambridge University Press (2000)
4. Csurka, G., Dance, C.-R., Fan, L., Willamowski, J., Bray, C.: Visual Categorization with Bag of Keypoints. In: Proc. of European Conference on Computer Vision 2004 Workshop on Statistical Learning in Computer Vision, pp. 59–74 (2004)
5. Dalal, N., Triggs, B.: Histogram of oriented gradients for human detection. In: Proc. of CVPR 2005 (2005)
6. Hariharan, B., Malik, J., Ramanan, D.: Discriminative decorrelation for clustering and classification. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part IV. LNCS, vol. 7575, pp. 459–472. Springer, Heidelberg (2012)
7. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning -Data Mining, Inference, and Prediction, 2nd edn. Springer (2006)
8. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines (2001), http://www.csie.ntu.edu.tw/=sjlin/libsvm
9. Mika, S., Rätsch, G., Weston, J., Schölkopf, B., Smola, A., Müller, K.: Fisher discriminant analysis with kernels. In: Proc. IEEE Neural Networks for Signal Processing Workshop, pp. 41–48 (1999)
10. Nishida, K., Kurita, T.: Kernel Feature Selection to Improve Generalization Performance of Boosting Classifiers. In: The 2006 International Conference on Image Processing, Computer Vision, & Pattern Recognition, Monte Carlo Resort, Las Vegas, Nevada, June 26-29 (2006)
11. Nishida, K., Kurita, T.: RANSAC-SVM for Large-Scale Datasets. In: Proc. of International Conference on Pattern Recognition, December 8-11. Tampa Convention Center, Tampa (2008)
12. Scholköpf, B., Burges, C.-J.-C., Smola, A.-J.: Advances in Kernel Methods - Support Vector Learning. The MIT Press (1999)
13. Schölkopf, B., Mika, S., Burges, C.-J.-C., Knirsch, P., Müller, K.-R., Rätsch, G., Smola, A.-J.: Input Spcae Versus Feature Space in Kernel-Based Methods. IEEE Trans. on Neural Networks 10(5), 1000–1017 (1999)
14. Suykens, J.-A.-K., Gestel, T.V., Vandewalle, J., De Moor, B.: A Support Vector Machine Formulation to PCA Analysis and Its Kernel Version. IEEE Trans. on Neural Networks 14(2), 447–450 (2003)
15. Vapnik, V.-N.: Statistical Learning Theory. John Wiley & Sons (1998)
16. Yan, S., Xu, D., Zhang, B., Zhang, H.-J.: Graph embedding: a general framework for dimensionality reduction. In: Proc. of CVPR 2005 (2005)