Multiple Kernel Learning Method Using MRMR Criterion and Kernel Alignment

Peng Wu[1,2], Fuqing Duan[1], and Ping Guo[1,⋆]

[1] Image Processing and Pattern Recognition Laboratory
Beijing Normal University
Beijing 100875, China
[2] Shandong Provincial Key Laboratory of Network Based Intelligent Computing
University of Jinan
Jinan 250022, China
pengwiseujn@gmail.com, fqduan@bnu.edu.cn, pguo@ieee.org

Abstract. Multiple kernel learning (MKL) is a widely used kernel learning method, but how to select kernel is lack of theoretical guidance. The performance of MKL is depend on the users' experience, which is difficult to choose the proper kernels in practical applications. In this paper, we propose a MKL method based on minimal redundant maximal relevance criterion and kernel alignment. The main feature of this method compared to others in the literature is that the selection of kernels is considered as a feature selection issue in the Hilbert space, and can obtain a set of base kernels with the highest relevance to the target task and the minimal redundancies among themselves. Experimental results on several benchmark classification data sets show that our proposed method can enhance the performance of MKL.

Keywords: minimal redundant maximal relevance, kernel alignment, kernel selection, multiple kernel learning.

1 Introduction

Multiple kernel learning (MKL) is an important kernel method, in which the most attractive character is so called 'kernel trick". MKL has been a hot research spot due to its success in lots of fields, such as bioinformatics [1], computer vision [2] and natural language processing [3]. MKL can be effortlessly derived from the canonical kernel method, i.e., support vector machine (SVM) [4]. Compared to SVM, MKL has a higher performance because of using a linear or nonlinear combination of several base kernels instead of only one specific kernel. Consequently, MKL aims at learning the combination coefficient of base kernels and some other parameters which are also learned by SVM. Lanckriet et al. [5] formulated it as a semi-definite programming problem. Bach et al. [6] reformulated it a quadratically constrained quadratic programming problem. Sonnenburg et al. [1] treated it as a second order cone programming problem that can be efficiently solved

⋆ Corresponding author.

M. Lee et al. (Eds.): ICONIP 2013, Part I, LNCS 8226, pp. 113–120, 2013.

using interior point methods and Rakotomanon et al. [7] addressed it through a weighted 2-norm regularization formulation with an additional constraint on the weights that encourage sparse kernel combination. Recently, localized MKL proposed by Gonen et al. [8], and the two-stage techniques for learning kernels based on a notion of alignment for MKL reported in [9] are two representatives methods.

Almost exclusively, methods aforementioned leave the task of selecting base kernels to users. It would be difficult in practice to choose a set of appropriate base kernels without prior knowledge, which maybe degrade the performance of MKL. To alleviate the negative effects, one can produce as many as possible candidate kernels, e.g., a family of polynomial kernels of arbitrary degree or a family of Gaussian kernels with different variances restricted in a specific range, and use all of them directly as base kernels. However, base kernels selected like that contain much redundant information and will give rise to high computation cost. Alternatively, one can only choose partial kernels with the highest relevance to the target task. Actually, to select a set of base kernels from a prescribed set of candidate kernels can be treat as a feature selection problem within the Hilbert feature space. Feature selection methods allow obtaining shorter training time and enhanced generalization by reducing over-fitting when constructing predictive models [10]. One state-of-the-art feature selection method, i.e., minimal redundancy maximal relevance (MRMR) [11], can be used as a filter in order to obtain a minimal subset of candidate kernels by reducing the redundancies among the selected kernels to a minimum. In this paper, we propose a MKL method based on the combination of MRMR, which is used as a filter, and kernel alignment, which is used to measure the mutual dependence between candidate kernels and target kernel, to select base kernels to enhance the performance of MKL. Note that, kernel alignment has been used for leaning a combination kernel from a prescribed candidate kernels, see in [9] [12]. Contrast to the previous work, in this study we take kernel alignment to select a set of base kernels instead of a combination kernel, which leads us to be more flexible in choosing the final combination form (linear, nonlinear or data-dependent) of base kernels.

The remainder of this paper is organized as follows: Section 2 reviews MRMR and kernel alignment. In Section 3, we describe the proposed method in detail. Experimental results on several benchmark classification data sets are reported and analyzed in Section 4, and our conclusions and further work are presented in the last section.

2 Minimal Redundancy Maximal Relevance and Kernel Alignment

2.1 Minimal Redundancy Maximal Relevance

MRMR is a well-known feature selection method based on the maximal statistical dependence of the target class on the data distribution. The mutual

information is a quantity that measures the mutual dependence of two random variables. Given variables $\mathbf{x}$ and $\mathbf{y}$, the mutual information between them can be calculated as follows

$$I(\mathbf{x};\mathbf{y}) = \sum_{x\in\mathbf{X}}\sum_{y\in\mathbf{y}} \log\frac{p(x;y)}{p(x)p(y)}, \tag{1}$$

where $p(x;y)$, $p(x)$ and $p(y)$ denote the joint probability distribution function of $\mathbf{x}$ and $\mathbf{y}$, the marginal probability distribution function of $\mathbf{x}$ and the marginal probability distribution function of $\mathbf{y}$, respectively. The higher value of $I(\mathbf{x};\mathbf{y})$ indicates the more mutual information they share, i.e., $\mathbf{x}$ is more correlated to $\mathbf{y}$.

Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_i\}$ be the whole feature set of a given data set, and $\mathbf{S}_m$, consisting of m features, be a selected subset of $\mathbf{X}$. Given $\mathbf{c}$, which represents the target class label, and $\mathbf{x}_i$, which represents a feature, we can obtain $\mathbf{S}_m$ by selecting the top m features in the descent order of $I(\mathbf{x}_i;\mathbf{c})$, but it is not a good scheme because of its failure in reducing the redundancy between the selected features. MRMR can select features that have the highest relevance to $\mathbf{c}$ and are also minimally redundant. In the algorithm of MRMR, first, all mutual information between candidate features and target class are calculated, and next, the mean mutual information between candidate features and the selected feature in subset $\mathbf{S}_{m-1}$, which has $m-1$ selected features, are calculated, and then to select the m_{th} feature from set $\{\mathbf{X}-\mathbf{S}_{m-1}\}$ according to the condition shown as follows:

$$\max_{\mathbf{x}_j\in\mathbf{X}-\mathbf{S}_{m-1}} [I(\mathbf{x}_j;\mathbf{c}) - \frac{1}{m-1}\sum_{\mathbf{x}_i\in\mathbf{S}_{m-1}} I(\mathbf{x}_j;\mathbf{x}_i)], \tag{2}$$

2.2 Kernel Alignment

In this paper, we use MRMR as a filter to select a set of base kernels from candidate kernels. The mutual information between two kernels can be calculated using kernel alignment which proposed by Cristianini et al. in [13]. Kernel alignment is a method to measure the similarity of two kernel matrices. Given a binary-class data set $\mathbf{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where y_i is the class label and $y_i \in \{-1, 1\}$, and $\mathbf{N}$ is the total number of samples, then the similarity between two kernel matrix on data set $\mathbf{S}$ is calculate by

$$A(\mathbf{S}, \mathbf{K}_1, \mathbf{K}_2) = \frac{\langle \mathbf{K}_1, \mathbf{K}_2\rangle_F}{\sqrt{\langle \mathbf{K}_1, \mathbf{K}_1\rangle_F \langle \mathbf{K}_2, \mathbf{K}_2\rangle_F}}, \tag{3}$$

where $\langle \mathbf{K}_1, \mathbf{K}_1\rangle_F$ is the inner product between kernel matrices, and the form is as follow

$$\langle \mathbf{K}_1, \mathbf{K}_2\rangle_F = \sum_{i,j=1}^{N} \mathbf{K}_1(x_i, x_j)\mathbf{K}_2(x_i, x_j). \tag{4}$$

If we consider $\mathbf{K}_2 = \mathbf{y}\mathbf{y}^T$, where $\mathbf{y}$ is the class vector of all samples, then we get

$$A(\mathbf{S}, \mathbf{K}_1, \mathbf{y}\mathbf{y}^T) = \frac{\langle \mathbf{K}_1, \mathbf{y}\mathbf{y}^T \rangle_F}{\sqrt{\langle \mathbf{K}_1, \mathbf{K}_1 \rangle_F \langle \mathbf{y}\mathbf{y}^T, \mathbf{y}\mathbf{y}^T \rangle_F}}. \tag{5}$$

In the next section, we will describe the proposed method in details.

3 The Proposed Method

The proposed method reported in this paper is based on a hybrid approach combining MRMR and alignment kernel (MRMRKA). The general algorithm of MRMRKA includes three steps: 1) some candidate kernels, or called candidate matrices, are generated based on the given data set, and 2) a set of base kernels is automatically selected by utilizing MRMRKA, and 3) the selected base kernels are fed into the process of MKL. The detailed steps are described as follows.

First, we obtain a set of candidate kernels. Kernel matrices are generated by the mapping of kernel functions. It is necessary to try to make use of several different kernel functions for getting some valid candidate kernels. Using linear kernel function, we generate the first kernel matrix, and the others can be generated by utilizing a family of polynomial kernel functions with different settings of the degree and a family of Gaussian kernel functions with variances in a prescribed interval.

Second, we select base kernels using MRMRKA. Given that the number of base kernels to be selected is m, the mutual information between candidate kernel matrices and target kernel matrix are calculated using (4), then select the candidate kernel with the maximal value of mutual information as the element of set $\mathbf{S}_1$. Then, use (2) or (3) to select the rest set $\mathbf{S}_i (2 \leq i \leq m)$. Note that, $I(\mathbf{x}_i; \mathbf{c})$ and $I(\mathbf{x}_i; \mathbf{x}_j)$ in (2) are substituted by (4) and (6), respectively.

Third, we execute MKL with the selected base kernels. Several MKL schemes can be chosen in this stage such as MKL based on semi-definite, MKL based on quadratically constrained quadratic programming, simpleMKL, localized MKL and so on.

4 Experiments

In this section, several experiments are done to test the proposed method on a number of classification data sets, and the experimental results of three different schemes to select base kernels in MKL are reported.

4.1 Data Sets and Preprocessing

Ten classification data sets which are available on the UCI machine learning archive [14] are adopted in the experiments, the detail informations about those data sets are shown in Table 1.

Table 1. Data sets information

Dataset	#Classes	#Attributes	#Instances
Blood	2	5	748
Breast	2	32	569
Control	6	60	600
Ecoil	5	8	336
Glass	7	10	214
Iris	3	5	150
Parkinsons	2	23	197
Seeds	3	7	210
Sonar	2	60	208
Wine	3	13	178

All raw data were preprocessed to have zero mean-value and unit variance. Each data set was divided randomly to three subsets with preserved class ratios. One of the three subsets was reserved as the testing set, and one of the remaining two was used as the training set and the other was used as the validation set. The validation sets of all data sets were used to optimize the parameter C, i.e., the trade-off parameter between model simplicity and classification error, by trying values $\{0.01, 0.1, 1, 10, 100\}$. The best C, i.e., leading to the highest classification accuracy on the validation set, was used to train the final classifier on the training set and its performance was measured over the testing set. The MKL scheme used in this study is LMKL due to its outstanding performance and we modified it to fit for multi-class classification tasks[1]. We repeated the experiment three times on each data set and reported the average classification accuracy as well as standard deviation.

4.2 Experimental Results and Comparison

For comparison, we adopt three schemes to select base kernels after producing candidate kernels: using all candidate kernels as base kernels (Method 1), selecting the top m kernels in the descent order of $A(\mathbf{S}, \mathbf{K}, \mathbf{y}\mathbf{y}^T)$ (Method 2), and selecting base kernels using MRMRKA. We produced forty candidate kernels on each data set, which consist of one linear kernel, four polynomial kernels with different degree values $\{2, 3, 4, 5\}$, and thirty-five Gaussian kernels with different variances whose values are limited in $[0.01, 1000]$.

We compared the performance of three schemes in terms of both computational time cost and classification accuracy, based on experiments on a quad-core 2.67G Xeon CPU running Windows 7 with the Matlab implementation. Table 2 shows the final results on all data sets. Note that, the number of selected base kernels, i.e., m, ranges from 2 to 16 in our experiments, so the results of Method 1 and Method 2 reported in Table 2 are the best classification accuracy. As we

[1] The original codes are available on `http://user.ics.aalto.fi/gonen/icml08.php`

Table 2. Classification results on all data sets

Dataset	Method 1		Method 2		MRMRKA	
	Accuracy	Time (s)	Accuracy	Time (s)	Accuracy	Time (s)
Blood	0.7590±0.0678	56±6.0	0.7711±0.0229	6±2.0	0.8005±0.0126	25±1.0
Breast	0.7964±0.2506	134±9.0	0.9613±0.0053	7±2.0	0.9754±0.0011	38±1.0
Control	0.9663±0.0816	1304±58	0.9680±0.0103	28±2.0	0.9731±0.0064	86±1.0
Ecoil	0.8142±0.0196	164±19	0.8443±0.0243	8±0.10	0.8682±0.0054	8±0.50
Glass	0.5728±0.0518	141±18	0.6122±0.0641	17±2.0	0.6352±0.0093	36±2.0
Iris	0.9804±0.0227	8±0.10	0.9804±0.0227	0.3±0.10	0.9804±0.0123	8±0.20
Parkinsons	0.8513±0.0335	21±8.0	0.8872±0.0506	6±2.0	0.9128±0.0196	7±2.0
Seeds	0.9275±0.0421	22±3.0	0.9372±0.0391	0.5±0.10	0.9420±0.0229	3±0.10
Sonar	0.5894±0.1034	108±29	0.5990±0.0549	0.6±0.10	0.6135±0.0474	6±0.50
Wine	0.9498±0.0269	48±3.0	0.9722±0.0170	2±0.40	0.9722±0.0092	13±0.20

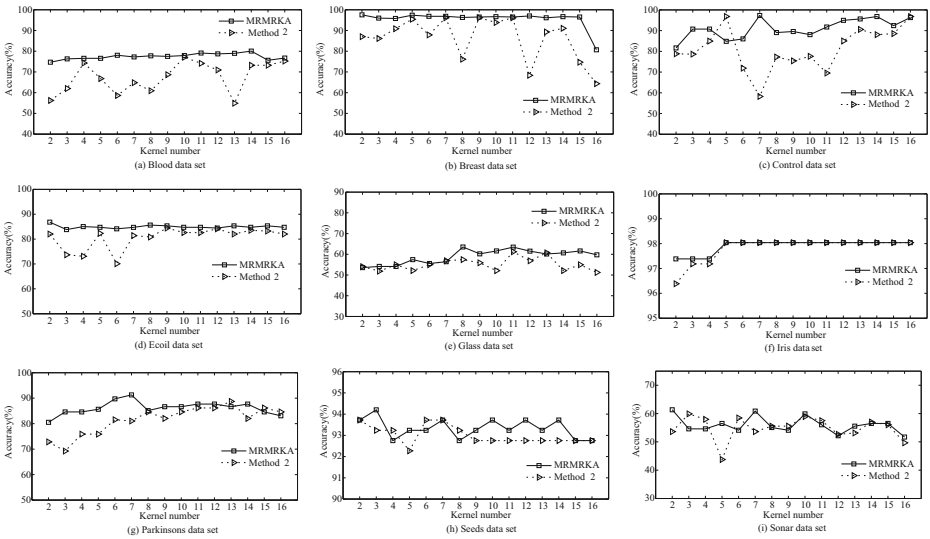

Fig. 1. Comparison of base kernels selected by MRMRKA and Method 2

can see from Table 2, MRMRKA obtains the best classification accuracy on all data sets. Both Method 2 and MRMRKA outperform Method 1 on all data sets. In addition, the time cost of Method 1 is higher than those of Method 2 and MRMRKA, because Method 1 does not take into count the redundancies among the selected base kernels and whether they are related to the target kernel, which undoubtedly has a negative effect on the efficiency and effectiveness of MKL. In the case of Method 2, the part of candidate kernels that are barely related to the target kernel is filtered in the selecting stage, but it does not take any measures to reduce the redundancies. Compared to Method 1 and Method 2, MRMRKA

selects the candidate kernels with the highest relevance to the target kernel and minimizes the redundancies among them, which makes it outperform the others. The time cost of MRMRKA is higher than that of Method 2 due to its a little bit more expensive, but the difference is acceptable. In general, Table 2 indicates MRMRKA gets a trade-off of the high accuracy and time cost.

We also examine that the number of selected kernels how to influence the performance of Method 2 and MRMRKA, and the results are shown in Fig. 1. The results demonstrate that the classification accuracy of MRMRKA fluctuates slightly with the change of m, in other words, when the number of selected base kernels is limited in a specific range, e.g., $[2, 16]$, the performance of MRMRKA are more stable than Method 2 in general. An important inspiration that can be drawn from the phenomenon is we can use cross-validation to select an optimal m by trying several finite values in practice. On the other hand, the fluctuation indicates that MRMRKA cannot entirely avoid the redundancies among the selected base kernels. To solve this kind problem, cross-validation can still be considered.

5 Conclusion and the Future Work

In this paper, in order to solve the problem of selecting base kernels in MKL, we propose a method which combine with MRMR and kernel alignment. The current results show that the proposed method can obtain a set of base kernels which can enhance the performance of MKL. There are two issues worth of further consideration. The first one is that only some medium-sized data sets are chosen in this study due to the computational time and space such kernels take when facing with large number of data samples, and the large scale data sets should be considered in the future work. The second one is that some other similarity measurements, such as Euclidean distance or Kullback-Leibler divergence can be utilized to measure the mutual dependence between kernels.

Acknowledgment. The research work in this paper was supported by the grants from the National Natural Science Foundation of China (Project No. 90820010, 61375045).

References

1. Sonnenburg, S., Rätsch, G., Schäfer, C., Schölkopf, B.: Large scale multiple kernel learning. The Journal of Machine Learning Research 7, 1531–1565 (2006)
2. Duan, L., Tsang, I.W., Xu, D., Maybank, S.J.: Domain transfer svm for video concept detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1375–1381. IEEE (2009)
3. Qi, M., Tsang, I.W.: Efficient Multi-template Learning for Structured Prediction. IEEE Transactions on Neural Networks and Learning Systems 24(2), 248–261 (2013)
4. Vapnik, V.: The nature of statistical learning theory. Springer (1999)

5. Lanckriet, G.R., Cristianini, N., Bartlett, P., et al.: Learning the kernel matrix with semidefinite programming. The Journal of Machine Learning Research 5, 27–72 (2004)
6. Bach, F.R., Lanckriet, G.R., Jordan, M.I.: Multiple kernel learning, conic duality, and the smo algorithm. In: Proceedings of the Twenty-First International Conference on Machine Learning, pp. 41–48. ACM (2004)
7. Rakotomamonjy, A., Bach, F., Canu, S., et al.: Simplemkl. Journal of Machine Learning Research 9, 2491–2521 (2008)
8. Gönen, M., Alpaydin, E.: Loclized algorithms for multiple kernel learning. Pattern Recognition 46, 795–807 (2013)
9. Cortes, C., Mohri, M., Rostamizadeh, A.: Algorithms for learning kernels based on centered alignment. The Journal of Machine Learning Research 13, 795–828 (2012)
10. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. The Journal of Machine Learning Research 3, 1157–1182 (2003)
11. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(8), 1226–1238 (2005)
12. Afkanpour, A., Szepesvari, C., Bowling, M.: Alignment based kernel learning with a continuous set of base kernels. arXiv preprint arXiv:1112.4607 (2011)
13. Cristianini, N., Shawe-Taylor, J., Elisseeff, A., et al.: On kernel-target alignment. In: NIPS, pp. 367–373 (2001)
14. Bache, K., Lichman, M.: UCI machine learning repository (2013)