Training of support vector machines with Mahalanobis kernels

Abe, Shigeo

Training of Support Vector Machines with Mahalanobis Kernels

Shigeo Abe

Graduate School of Science and Technology
Kobe University
Rokkodai, Nada, Kobe, Japan
abe@eedept.kobe-u.ac.jp
http://www2.eedept.kobe-u.ac.jp/~abe

Abstract. Radial basis function (RBF) kernels are widely used for support vector machines. But for model selection, we need to optimize the kernel parameter and the margin parameter by time-consuming cross validation. To solve this problem, in this paper we propose using Mahalanobis kernels, which are generalized RBF kernels. We determine the covariance matrix for the Mahalanobis kernel using the training data corresponding to the associated classes. Model selection is done by line search. Namely, first the margin parameter is optimized and then the Mahalanobis kernel parameter is optimized. According to the computer experiments for two-class problems, a Mahalanobis kernel with a diagonal covariance matrix shows better generalization ability than a Mahalanobis kernel with a full covariance matrix, and a Mahalanobis kernel optimized by line search shows comparable performance with that with an RBF kernel optimized by grid search.

1 Introduction

Support vector machines have been used for various applications as a powerful tool for pattern classification. One of the advantages of support vector machines is that we can improve generalization ability by proper selection of kernels. In most cases polynomial kernels and radial basis function network (RBF) kernels are used. Mahalanobis kernels [1], which exploit the data distribution information more than RBF kernels do, are expected to ease model selection but how to set the covariance matrix is a difficult problem. Friedrichs and Igel [2] used evolution strategies to tune the parameters obtained by grid search but it is time consuming.

In this paper, we propose model selection for Mahalanobis kernels. Namely, using the data belonging to the two classes, we calculate the covariance matrix for the Mahalanobis kernel. We then optimize the margin parameter and the kernel parameter that scales the Mahalanobis distance by line search: after optimizing the margin parameter by cross validation, we optimize the kernel parameter. We show the usefulness of Mahalanobis kernels over RBF kernels using two-class data sets.

In Section 2, we discuss Mahalanobis kernels, and in Section 3 we discuss model selection. Finally in Section 4, we compare performance of Mahalanobis kernels with RBF kernels.

2 Mahalanobis Kernels

First we explain the Mahalanobis distance between a datum and the center vector of a cluster. Let the set of M m-dimensional data be $\{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$ for the cluster. Then the center vector and the covariance matrix of the data are given, respectively, by

$$\mathbf{c} = \frac{1}{M}\sum_{i=1}^{M} \mathbf{x}_i, \tag{1}$$

$$Q = \frac{1}{M}\sum_{i=1}^{M} (\mathbf{x}_i - \mathbf{c})\,(\mathbf{x}_i - \mathbf{c})^T. \tag{2}$$

The Mahalanobis distance of $\mathbf{x}$ is given by

$$d(\mathbf{x}) = \sqrt{(\mathbf{x} - \mathbf{c})^T Q^{-1}\,(\mathbf{x} - \mathbf{c})}. \tag{3}$$

Because the Mahalanobis distance is normalized by the covariance matrix, it is linear translation invariant [3]. This is especially important because we need not worry about the scales of input variables.

Another interesting characteristic is that the average of the square of Mahalanobis distances is m [3]:

$$\frac{1}{M}\sum_{i=1}^{M} (\mathbf{x}_i - \mathbf{c})^T Q^{-1}\,(\mathbf{x}_i - \mathbf{c}) = m. \tag{4}$$

Based on the definition of the Mahalanobis distance, we define the Mahalanobis kernel by

$$H(\mathbf{x}, \mathbf{x}') = \exp\left(-(\mathbf{x} - \mathbf{x}')^T A\,(\mathbf{x} - \mathbf{x}')\right), \tag{5}$$

where A is a positive definite matrix. Here, the Mahalanobis distance is calculated between $\mathbf{x}$ and $\mathbf{x}'$, not between $\mathbf{x}$ and $\mathbf{c}$. The Mahalanobis kernel is an extension of the RBF kernel. Namely, by setting

$$A = \gamma\, I, \tag{6}$$

where $\gamma (> 0)$ is a parameter for slope control and I is the $m \times m$ unit matrix, we obtain the RBF kernel:

$$\exp(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2). \tag{7}$$

For a two-class problem, the Mahalanobis kernel is used for the data belonging to one of the two classes. Assuming that $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$ is the set of data belonging to one of the two classes, we calculate the center and the covariance matrix by (1) and (2), respectively.

Then we approximate the Mahalanobis kernel by

$$H(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\delta}{m}\,(\mathbf{x} - \mathbf{x}')^T Q^{-1}\,(\mathbf{x} - \mathbf{x}')\right), \tag{8}$$

where $\delta\,(> 0)$ is the scaling factor to control the Mahalanobis distance.

From (4), by dividing the square of the Mahalanobis distance by m, it is normalized to 1 irrespective of the number of input variables. Although (8) is an approximation of the Mahalanobis kernel, this may enable to limit the search of the optimal δ value in a small range.

If we use the full covariance matrix, it will be time-consuming for a large number of input variables. Thus we consider two cases: Mahalanobis kernels with diagonal covariance matrices and Mahalanobis kernels with full covariance matrices. Hereafter we call the former diagonal Mahalanobis kernels and the latter non-diagonal Mahalanobis kernels.

3 Model Selection

To maximize the generalization ability of the support vector machine we need to optimize the parameters by model selection. The most reliable method is cross validation. In the following, we discuss model selection for RBF kernels and Mahalanobis kernels by cross validation.

3.1 RBF Kernels

For RBF kernels, we need to determine the values of γ and C by grid search. To set the proper search range of γ, it is better to normalize the input ranges into $[0, 1]$. Thus, because the maximum value of $\|\mathbf{x} - \mathbf{x}'\|^2$ is m, we use the following RBF kernels instead of (7) [4]:

$$\exp\left(-\frac{\gamma}{m}\|\mathbf{x} - \mathbf{x}'\|^2\right). \tag{9}$$

However, because RBF kernels are not scale invariant, the range of $[0, 1]$ may not be optimal.

3.2 Mahalanobis Kernels

For Mahalanobis kernels, we need to determine the values of δ and C. But because Mahalanobis kernels given by (8) are determined according to the data distribution and normalized by m, the initial value of $\delta = 1$ is a good selection. Thus, we can carry out model selection by line search not by grid search. Namely, the model selection is done as follows:

1. Set $\delta = 1$ and determine the value of C by cross validation. We call this the first stage.
2. Setting the value of C as that determined by the first stage, determine the value of δ by cross validation. We call this the second stage.

Because $\delta = 1$ is a good initial value, we may search the optimal value around 1, e.g., $[0.1, 2]$.

In addition, because Mahalanobis kernels are normalized by the covariance matrix, it is scale invariant. Therefore, the scale transformation of input variables does not affect the classification performance of the support vector machine.

4 Performance Evaluation

We compared the generalization ability of Mahalanobis kernels and RBF kernels using two-class data sets used in [5].[1] Each problem has 100 or 20 training data sets and their corresponding test data sets. Because there is not much difference of generalization abilities between L1 and L2 support vector machines, we used L1 support vector machines. We determined the optimal values of γ and C for RBF kernels and those of δ and C for Mahalanobis kernels by 5-fold cross validation. Because the input ranges of the data sets were not normalized, we normalized them to $[0, 1]$

For RBF kernels for a value of γ in $\{0.1, 0.5, 1, 5, 10, 15\}$ we performed cross validation of the first five training data sets changing $C = [1, 10, 50, 100, 500, 1000, 2000, 3000, 5000, 8000, 10000, 50000, 100000]$, selected the optimal γ that showed the minimum average error rate for the five validation data sets, and selected the median of the best value of C for the optimal γ. Then, for the optimal values of γ and C, we trained the support vector machine for 100 or 20 training data sets and calculated the average recognition error and the standard deviation for the test data sets.

Similarly for Mahalanobis kernels, at the first stage we determined the optimal value of C by cross validation for the first five training data sets. Then, at the second stage we performed cross validation with the determined value of C, changing $\delta = [0.1, 0.2, \ldots, 1.9, 2]$. As a reference we also performed the grid search of optimum δ for $\delta = [0.1, 0.5, 1.0, 1.5, 2.0]$ and C.

If the recognition rate of the validation set took the maximum value for different values of C, we took the smallest value as the optimal value.

Table 1 lists the parameters obtained by the preceding procedure. Here, we do not include parameters for Mahalanobis kernels obtained by grid search. From the table, it is seen that the values of C for the Mahalanobis kernels are equal to or smaller than those for RBF kernels. In addition, for the image and thyroid data sets, the values for non-diagonal Mahalanobis kernels are smaller than for the diagonal Mahalanobis kernels. This means that the support vector machines with RBF kernels are the most difficult to fit to the data, whereas those with non-diagonal Mahalanobis kernels are the easiest.

[1] http://ida.first.fraunhofer.de/projects/bench/benchmarks.htm

Table 1. Parameter setting

Data	RBF		Diagonal		Non-diagonal	
	γ	C	C	δ	C	δ
Banana	15	100	50	0.8	50	0.9
B. Cancer	1	10	1	0.6	1	0.8
Diabetes	10	1	1	0.5	1	0.2
German	5	1	1	1.7	1	0.9
Heart	0.1	50	1	0.2	1	0.1
Image	10	1000	500	0.7	100	1
Ringnorm	15	1	1	1.5	1	1.3
F. Solar	1	1	1	0.1	1	0.1
Splice	10	10	10	0.8	10	0.5
Thyroid	5	1000	50	0.4	10	0.9
Titanic	10	10	10	0.7	10	0.6
Twonorm	1	1	1	0.9	1	0.2
Waveform	5	10	1	0.6	1	0.4

Table 2 lists the average classification errors and the standard deviations with the ± symbol. The "Diagonal-1" and "Diagonal-2" columns list the values for the first and second stages, respectively, and the "Diagonal" column lists the values by the grid search. Performance of RBF kernels with the input range of $[0, 1]$ is different from that with the original input range given in [5]. Except for the ringnorm data set, the performance with the input range of $[0, 1]$ performed better. If we use the original input range for the ringnorm data set, the performance is 1.7±0.1, which is equivalent to that of the second stage using the diagonal Mahalanobis kernel (Diagonal-2). But for Mahalanobis kernels, performance does not change for the change of the input range.

The best performance in the row is shown in boldface. Except for the diabetes, heart, and f. solar data sets, the recognition performance of diagonal Mahalanobis kernels with $\delta = 1$ (Diagonal-1) was comparable with that of the RBF kernels. For these data sets by optimizing the value of δ, performance of the diagonal Mahalanobis kernels (Diagonal-2) was improved and comparable with that of RBF kernels. There is not much difference between Diagonal-2 and Diagonal. But performance of non-diagonal Mahalanobis kernels was not so good. The full covariance matrix might cause overfitting.

Table 2. Comparison of average error rates and standard deviations.

Data	RBF	Diagonal-1	Diagonal-2	Diagonal	Non-diagonal
Banana	10.5±0.5	10.5±0.4	**10.4**±0.5	**10.4**±0.5	10.5±0.5
B. Cancer	**25.6**±4.4	25.9±4.2	**25.6**±4.4	25.9±4.2	26.1±4.4
Diabetes	23.4±1.7	24.7±1.9	23.7±1.7	23.7±1.7	**23.3**±1.8
German	23.8±2.1	**23.4**±2.1	23.9±2.1	23.7±1.7	23.7±2.2
Heart	16.1±3.1	17.2±3.2	15.7±3.2	**15.6**±3.4	17.2±4.0
Image	**2.8±0.5**	3.1±0.6	3.0±0.5	3.0±0.6	3.2±0.6
Ringnorm	2.6±0.4	1.8±0.2	1.7±0.1	**1.6±0.1**	1.8±0.1
F. Solar	**32.3**±1.8	34.1±2.0	32.5±1.7	32.8±1.7	32.5±1.7
Splice	10.8±0.7	**10.7**±0.7	10.8±0.6	10.8±0.7	13.0±0.6
Thyroid	**4.1**±2.3	4.2±2.0	**4.1**±2.3	4.2±2.3	6.9±2.8
Titanic	**22.5±1.0**	**22.5±1.0**	**22.5±1.0**	**22.5±1.0**	22.6±1.0
Twonorm	**2.4±0.1**	2.7±0.2	2.7±0.2	2.7±0.1	2.8±0.2
Waveform	10.3±0.4	**9.9±0.4**	**9.9**±0.5	10.5±0.4	15.6±1.2

5 Conclusions

We discussed how to train support vector machines with Mahalanobis kernels for pattern classification problems. We calculate the covariance matrix using the training data and determine the optimum values of the margin parameter and the kernel parameter by line search. The computer experiments showed that the performance of the Mahalanobis kernels by line search of the optimal margin and kernel parameters was comparable to that of RBF kernels by grid search of the optimal parameters.

References

1. R. Herbrich. *Learning Kernel Classifiers: Theory and Algorithms.* MIT Press, Cambridge, MA, 2002.
2. F. Friedrichs and C. Igel. Evolutionary tuning of multiple SVM parameters. *Proc. ESANN 2004*, pp. 519–524, 2004.
3. S. Abe. *Pattern Classification: Neuro-Fuzzy Methods and Their Comparison.* Springer-Verlag, London, 2001.
4. S. Abe. *Support Vector Machines for Pattern Classification.* Springer-Verlag, New York, 2005.
5. K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf. An introduction to kernel-based learning algorithms. *IEEE Trans. Neural Networks*, 12(2):181–201, 2001.