



ELSEVIER

Pattern Recognition Letters 20 (1999) 513–519

Pattern Recognition
Letters

www.elsevier.nl/locate/patrec

Subspace classifier in the Hilbert space

Koji Tsuda *

Machine Understanding Division, Electrotechnical Laboratory, 1-1-4, Umezono, Tsukuba-shi, Ibaraki-ken 305-8568, Japan

Received 7 October 1988; received in revised form 15 February 1999

Abstract

To improve the performance of the subspace classifier, it is effective to reduce the dimensionality of the intersections between subspaces. For this purpose, the feature space is mapped implicitly to the infinite dimensional Hilbert space and the subspace classifier is applied in the Hilbert space. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Pattern recognition; Subspace classifier; Hilbert space; Kernel functions; Support vector machine

1. Introduction

The subspace classifier is the pattern recognition method that uses a linear subspace for describing a class (Oja, 1983). In training, a subspace is fit to training samples so that the sum of the squared distances between the samples and the subspace is minimized. In classifying an unlabeled sample, the sample is classified to the class whose subspace is the nearest. The subspace classifier was developed in the early days of pattern recognition study, but nowadays it is still in focus (Deshuang, 1996; Prakesh and Murty, 1996, 1997).

It is known that the performance of the subspace classifier is affected by the dimensionality of the intersections of subspaces (Oja, 1983). When the dimensionality of the feature space is larger than the number of training samples, the subspace spanned by the training samples of a particular class have no intersection with the other sub-

spaces. However, since the number of training samples is much greater than the dimensionality of the feature space in many cases, the intersection of subspaces cannot be avoided. When an unlabeled sample is placed in the intersection of two subspaces, the distance between the sample and each subspace is zero, and so this sample cannot be classified. In neighborhood areas of the intersection, the difference between the distances to the two subspaces is so small that the classification becomes doubtful. Therefore, the dimensionality of the intersections should be small as possible.

Usually, to keep the dimensionality of the intersections small, the dimensionality of each subspace is reduced using the principal component analysis (Oja, 1983). However, on the contrary, by increasing the dimensionality of the feature space, the dimensionality of the intersections can be reduced. In this paper, our aim is to improve the performance of the subspace classifier by mapping the feature space to a higher dimensional space. For this purpose, we utilize the implicit high dimensional casting method that is used in the support vector machine (SVM) (Vapnik, 1995).

*Tel.: +81-298-54-5912; fax: +81-298-54-5949; e-mail: tsudak@etl.go.jp

Let the p -dimensional feature space be described as $\mathfrak{R}^p$, and let φ be the continuous mapping from $\mathfrak{R}^p$ to a higher dimensional space $\mathfrak{R}^q$. It is assumed that the following equation holds for every $\mathbf{x}, \mathbf{y} \in \mathfrak{R}^p$:

$$\langle \varphi(\mathbf{x}), \varphi(\mathbf{y}) \rangle = K(\mathbf{x}, \mathbf{y}), \quad (1)$$

where K is a positive definite kernel function and $\langle \cdot, \cdot \rangle$ denotes the inner product in $\mathfrak{R}^q$. The most frequently used kernel function is the Gaussian function

$$K(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{\sigma^2}\right), \quad (2)$$

where σ is a width parameter. In SVM, the feature space is mapped to a higher dimensional space $\mathfrak{R}^q$ by φ and the classification is performed in $\mathfrak{R}^q$.

When the Gaussian kernel is used, the existence of the mapping φ is proved only when $q = \infty$ (Osuna et al., 1997). So, the image space of φ is the infinite dimensional Hilbert space l_2 (Small, 1994). It is clear that the mapping to an infinite dimensional space cannot directly be implemented on a computer. But, since the existence of the mapping is proved, the inner product between two points in the Hilbert space is obtained by Eq. (1). Based on the inner product, the classification can be performed without knowing the mapping explicitly. In fact, SVM constructs a linear classifier in the Hilbert space (Scholkopf et al., 1997).

We will show that the dimensionality of the intersections is always zero in the Hilbert space. Therefore, it is expected that the performance of the subspace classifier in the Hilbert space would be better than the one in the original feature space. In this paper, we call this classifier “the subspace classifier in the Hilbert space”. This classifier is actually applied to a Hiragana (i.e., Japanese character) recognition problem, and obtained high classification accuracy.

This paper is organized as follows. In Section 2, the conventional subspace classifier is introduced. In Section 3, we propose the subspace classifier in the Hilbert space. In Section 4, we will prove that the dimensionality of the intersection becomes zero in the Hilbert space. In Section 5, the Hiragana recognition experiment is performed to

compare our classifier with conventional classifiers, and the difference between our classifier and SVM is discussed. Section 6 is the conclusion.

2. Subspace classifier

In this section, we review the subspace classifier. Although there are many variants of the subspace classifier (Prakesh and Murty, 1996, 1997), we introduce the most fundamental one, that is, the CLAFIC method (Oja, 1983).

Let $C_1, \dots, C_c$ denote the classes and let the training samples of C_k be denoted as $\mathbf{s}_{ki}$ ($1 \leq i \leq n$). Let $\mathcal{D}_k$ be the r_k dimensional subspace of C_k whose orthonormal bases are $\mathbf{t}_{ki}$ ($1 \leq i \leq r_k$). Also, let T_k be the matrix whose column vectors are $\mathbf{t}_{ki}$ ($1 \leq i \leq r_k$).

Since $\mathcal{D}_k$ is contained in the subspace spanned by the training samples, the orthonormal bases can be rewritten as follows:

$$\mathbf{t}_{ki} = \sum_{j=1}^n u_{kij} \mathbf{s}_{kj} = S_k \mathbf{u}_{ki}, \quad (3)$$

where

$$S_k = (\mathbf{s}_{k1}, \dots, \mathbf{s}_{kn}), \quad (4)$$

$$\mathbf{u}_{ki} = (u_{ki1}, \dots, u_{kin})^T. \quad (5)$$

When the unlabeled sample $\mathbf{v}$ is classified by the subspace classifier, the distance between $\mathbf{v}$ and each subspace is calculated. Then, $\mathbf{v}$ is classified to the class with the smallest distance. The distance between $\mathbf{v}$ and $\mathcal{D}_k$ is described as follows:

$$d_k = \|\mathbf{v} - T_k T_k^T \mathbf{v}\|^2 \quad (6)$$

$$= \|\mathbf{v}\|^2 - \|T_k^T \mathbf{v}\|^2. \quad (7)$$

Since the first term is independent from the class, it can be omitted. Then, the membership function of C_k is as follows:

$$f_k(\mathbf{v}) = \|T_k^T \mathbf{v}\|^2. \quad (8)$$

Since the membership function describes the degree that $\mathbf{v}$ belongs to C_k , the unknown sample $\mathbf{v}$ is classified to the class with the largest $f_k(\mathbf{v})$. Substituting Eq. (3) into Eq. (8), we obtain

$$f_k(\mathbf{v}) = \|U_k^T S_k^T \mathbf{v}\|^2, \quad (9)$$

where U_k denotes the matrix whose column vectors are $\mathbf{u}_{ki}$ ($i = 1, \dots, n$). Note that $S_k^T \mathbf{v}$ is the vector of inner products between the unlabeled sample and the training samples.

In training the subspace classifier, the sum of the squared distances between the training samples and the subspace is minimized, which can also be achieved by maximizing the sum of the norms of the training samples projected on the subspace. The optimization problem for training is described as follows:

Find U_k that maximizes

$$\sum_{i=1}^n \|U_k^T S_k^T \mathbf{s}_{ki}\|^2 \quad (10)$$

subject to the constraints

$$U_k^T S_k^T S_k U_k = I,$$

where I is an $n \times n$ unit matrix. The optimal solution is given by the following theorem.

Theorem 1. Let the eigenvalues of the matrix $P_k = S_k^T S_k$ be denoted by λ_{ki} ($1 \leq i \leq n$) in descending order, and let $\mathbf{a}_{ki}$ denote the corresponding eigenvector of λ_{ki} . Then, the optimal solution of Eq. (10) is as follows:

$$U_k = A'_k (L'_k)^{-1}, \quad (11)$$

where

$$A'_k = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_{r_k}),$$

$$L'_k = \text{diag}(\sqrt{\lambda_{k1}}, \sqrt{\lambda_{k2}}, \dots, \sqrt{\lambda_{kr_k}}).$$

Proof. Let R denote the rank of P_k . The matrix P_k is decomposed as follows:

$$P_k = A_k L_k L_k^T A_k^T, \quad (12)$$

where L_k is described as

$$L_k = \text{diag}(\sqrt{\lambda_{k1}}, \dots, \sqrt{\lambda_{kR}}), \quad (13)$$

and A_k is the matrix whose column vectors are $\mathbf{a}_{k1}, \dots, \mathbf{a}_{kR}$.

Let $X = L_k^T A_k^T U_k$ and $\mathbf{y}_i = L_k^{-1} A_k^T S_k^T \mathbf{s}_{ki}$. Then, Eq. (10) can be rewritten as follows:

Find X that maximizes

$$\sum_{i=1}^n \mathbf{y}_i^T X X^T \mathbf{y}_i \quad (14)$$

subject to the constraint

$$X^T X = I.$$

The optimal solution of Eq. (14) can be obtained from the eigenvectors of the correlation matrix of $\mathbf{y}_i$ (Bellman, 1970). The correlation matrix is described as follows:

$$\sum_{i=1}^n \mathbf{y}_i \mathbf{y}_i^T = L_k^{-1} A_k^T P_k P_k^T A_k L_k^{-1} \quad (15)$$

$$= L_k^T L_k \quad (16)$$

$$= \text{diag}(\lambda_{k1}, \dots, \lambda_{kR}). \quad (17)$$

Since the correlation matrix is diagonal, the eigenvalues are $\lambda_{k1}, \dots, \lambda_{kR}$ in descending order, and the eigenvector that corresponds to λ_i is $\mathbf{e}_i$ which is the R -dimensional vector whose i th element is 1 and all the other elements are 0. The optimal X is described as follows:

$$X = L_k^T A_k^T U_k = (\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{r_k}). \quad (18)$$

The optimal U_k is derived from Eq. (18) as

$$U_k = A'_k (L'_k)^{-1}. \quad \square \quad (19)$$

According to the theorem, the training of the subspace classifier can be performed by the matrix computations to obtain eigenvalues and eigenvectors. Since such computations are included in most computational libraries, the subspace classifier is easy to implement.

It is known that the dimensionality of each subspace deeply affects the classification performance. It is better to change the dimensionality for each subspace than to fix the dimensionality over all classes. The criterion called “fidelity” is often used to determine the dimensionality (Oja, 1983), where the dimensionality is determined so that the fidelity of each subspace becomes the same pre-determined value. The fidelity of C_k is obtained as follows:

$$y_k(r_k) = \frac{1}{n} \sum_{i=1}^n \sqrt{\sum_{j=1}^{r_k} (\mathbf{u}_{kj}^T S_k^T \mathbf{s}_{ki})^2 / \|\mathbf{s}_{ki}\|^2}. \quad (20)$$

3. Subspace classifier in the Hilbert space

In this section, we consider the subspace classifier in the infinite dimensional real Hilbert space H . The elements of H are real infinite dimensional vectors

$$\mathbf{x} = (x_1, x_2, \dots)^T, \quad (21)$$

whose squared sum of the elements is finite, i.e.,

$$\sum_{i=1}^{\infty} x_i^2 < \infty. \quad (22)$$

When the feature space $\mathcal{R}^p$ is mapped to H by φ , the unknown sample $\mathbf{v}$ and the training sample $\mathbf{s}_{ki}$ are mapped to $\varphi(\mathbf{v})$ and $\varphi(\mathbf{s}_{ki})$, respectively. According to Eq. (1), we have the following equations:

$$\langle \varphi(\mathbf{v}), \varphi(\mathbf{s}_{ki}) \rangle = \exp \left(-\frac{\|\mathbf{v} - \mathbf{s}_{ki}\|^2}{\sigma^2} \right), \quad (23)$$

$$\langle \varphi(\mathbf{s}_{ki}), \varphi(\mathbf{s}_{kj}) \rangle = \exp \left(-\frac{\|\mathbf{s}_{ki} - \mathbf{s}_{kj}\|^2}{\sigma^2} \right). \quad (24)$$

In the Hilbert space, the same algorithm is performed as the conventional subspace classifier. So, in order to obtain the membership function in the Hilbert space, all you have to do is to replace the inner product in Eq. (9) by the one in the Hilbert space:

$$f_k(\varphi(\mathbf{v})) = \|U_k^T \mathbf{g}(\varphi(\mathbf{v}))\|^2, \quad (25)$$

where

$$\mathbf{g}(\varphi(\mathbf{v})) = (\varphi(\mathbf{v})^T \varphi(\mathbf{s}_{k1}), \dots, \varphi(\mathbf{v})^T \varphi(\mathbf{s}_{kn}))^T.$$

Also in training, the inner product in the optimization is replaced as follows:

Find U_k that maximizes

$$\sum_{i=1}^n \|U_k^T \mathbf{g}(\varphi(\mathbf{s}_{ki}))\|^2 \quad (26)$$

subject to the constraints

$$U_k^T G U_k = I, \quad (27)$$

where G is the matrix whose column vectors are $\mathbf{g}(\varphi(\mathbf{s}_{ki}))$ ($i = 1, \dots, n$). The optimal solution can be obtained in the same way as the conventional subspace classifier.

4. Discussion on the intersection of subspaces

The classification performance of the subspace classifier is said to be deeply affected by the dimensionality of the intersections of the subspaces. The dimensionality of the intersections should be small to improve the classification accuracy. In this section, we will show that the dimensionality of the intersection is always zero in the Hilbert space.

The intersection of the subspaces $\mathcal{D}_i$ and $\mathcal{D}_j$ is described as $\mathcal{D}_i \cap \mathcal{D}_j$. The intersection is also a linear subspace whose dimensionality is given as follows:

$$\begin{aligned} \dim(\mathcal{D}_i \cap \mathcal{D}_j) &= \dim(\mathcal{D}_i) + \dim(\mathcal{D}_j) \\ &\quad - \dim(\mathcal{D}_i \cup \mathcal{D}_j). \end{aligned} \quad (28)$$

First, we will prove the following theorem.

Theorem 2. *When a finite set of samples is mapped to the Hilbert space H by φ , all mapped samples are linearly independent.*

Proof. Let the mapped samples be described as $\varphi(\mathbf{s}_i)$ ($1 \leq i \leq n$). To show that these samples are linearly independent, it is sufficient to show that the Gram matrix M is regular (Bellman, 1970), where the (i, j) -element of M is described as follows:

$$m_{ij} = \langle \varphi(\mathbf{s}_i), \varphi(\mathbf{s}_j) \rangle = \exp \left(-\frac{\|\mathbf{s}_i - \mathbf{s}_j\|^2}{\sigma^2} \right). \quad (29)$$

The following theorem is known (Haykin, 1995).

Theorem 3. *Let $\mathbf{s}_1, \dots, \mathbf{s}_n$ be the points on $\mathcal{R}^p$, each of which is different from the other samples. Then, the $n \times n$ matrix whose (i, j) -element is described as*

$$\exp \left(-\frac{\|\mathbf{s}_i - \mathbf{s}_j\|^2}{\sigma^2} \right) \quad (30)$$

is positive definite.

By Theorem 3, it is clear that M is positive definite and accordingly regular.

By Theorem 2, the mapped training samples of C_i and C_j ($i \neq j$) are linearly independent in the Hilbert space. So, the mapped samples of each class span an n -dimensional subspace:

$$\dim(\mathcal{D}_i) = \dim(\mathcal{D}_j) = n. \quad (31)$$

Also, the training samples of both classes span a $2n$ -dimensional subspace:

$$\dim(\mathcal{D}_i \cup \mathcal{D}_j) = 2n. \quad (32)$$

Then, substituting Eqs. (31) and (32) into Eq. (28), we obtain the following equation:

$$\dim(\mathcal{D}_i \cup \mathcal{D}_j) = \dim(\mathcal{D}_i) + \dim(\mathcal{D}_j). \quad (33)$$

Therefore, it is shown that the dimensionality of the intersection is 0.

5. Hiragana recognition experiment

5.1. Settings

In this section, we will compare our classifier with the conventional subspace classifier, support vector machine and k -nearest neighbor in the handwritten Hiragana recognition problem with 71 classes. The character images were taken from the ETL9B database¹ which contains 200 images per a character. The samples are shown in Fig. 1. In feature extraction, the contour direction histogram feature (Kimura et al., 1987), whose dimensionality is 196, was used.

In a preliminary experiment, the optimal value of the width parameter σ is sought for our classifier and SVM, respectively. In this experiment, the number of training samples was 20 and the number of test samples was also 20. The value of σ was changed from 10 to 100 by the interval of 10. The fidelity value of our classifier was set to 0.99. As a result, our classifier achieved the smallest error rate when $\sigma = 50$, and SVM achieved when $\sigma = 30$. In the following experiment, these values are used for σ .

Next, our classifier is compared with the conventional subspace classifier. The number of training samples was 160 per class, and the number of testing samples was 40. The samples were chosen randomly and 10 trials were made with dif-

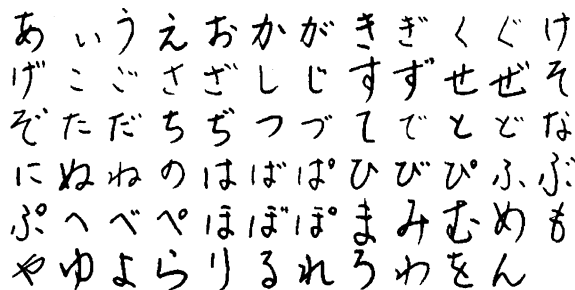


Fig. 1. Samples of handwritten Hiragana images (71 characters).

ferent choice of samples. The error rate described in the following is the average over 10 trials.

5.2. Results

Fig. 2 shows the error rates of the conventional subspace classifier and the subspace classifier in the Hilbert space when the fidelity changes from 0.95 to 0.995. The minimum error rate of our classifier was 3.64% when the fidelity is 0.99. On the other hand, the minimum error rate of the conventional classifier was 4.98% when the fidelity is 0.98. The error rate was reduced by 26.9%, so you can see that the error rate is improved significantly.

We also compared our classifier with SVM. The experimental setting is the same as in the previous

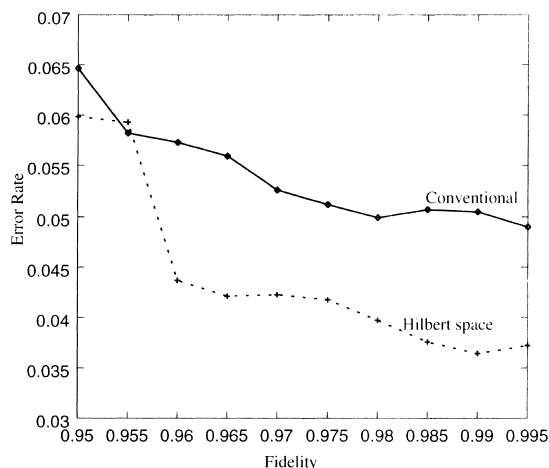


Fig. 2. The error rates of the conventional subspace classifier and the subspace classifier in the Hilbert space.

¹ For information, see <http://www.etl.go.jp/etlcdb>

Table 1

The error rates of the subspace classifier in the Hilbert space, the support vector machine and conventional classifiers

	Error rate (%)
Subspace classifier in the Hilbert space	3.64
Support vector machine	3.02
Subspace classifier	4.98
1-NN	8.73
3-NN	7.71
5-NN	7.35
7-NN	7.74

experiment. For each class, SVM is trained to learn the boundary between the class and all the other classes. The unlabeled sample is classified to the class for which SVM outputs the largest value.

The error rates of our classifier and SVM are shown in Table 1. This table also shows the error rates of the conventional subspace classifier and the k -nearest neighbor classifier. It shows that SVM achieved the smallest error rate 3.02%. The reason is considered that SVM is trained by the training samples of its own class and rival classes, whereas our classifier is trained by the samples of its own class. However, SVM has the problem that the training needs enormous computational cost, because the optimization problem becomes very large, which will be discussed later in the next section. In comparison with the conventional subspace classifier and k -nearest neighbor, our classifier and SVM are superior by far. This is the good effect of the mapping to the Hilbert space.

5.3. Comparison with support vector machine

In this section, we will compare our classifier with SVM from various viewpoints. The common features of the two classifiers are described as follows:

- The two classifiers utilize the mapping to the Hilbert space.
- The training is performed by solving a quadratic optimization problem, so the global optimal solution can be easily obtained.

The first point is the biggest feature of the two classifiers. In conventional ways of pattern recognition, the dimensionality of the feature space is usually reduced to obtain better results. On the contrary, the two classifiers use the mapping which increases the dimensionality.

The second point is desirable from a practical viewpoint. So far, most of the high performance classifiers use nonlinear discriminant function and often suffer from local minima problems. Although the two classifiers use nonlinear discriminant functions, the optimization problem is quadratic and local minima problems do not exist.

The difference between the two classifiers is described as follows:

- Whereas SVM forms a linear discrimination boundary in the Hilbert space, our classifier forms a quadratic boundary.
- SVM is trained by the training samples of its own class and rival classes, whereas our classifier is trained by the samples of its own class only.
- The size of the optimization problem for training is much larger in SVM.

From the first point, we can see that, even two classes are not linearly separable, our classifier may be able to discriminate the classes. But, in actual experiment, the error rate was inferior to SVM. It is because of the second point: whereas SVM is trained using the information from rival classes, our classifier ignores the information.

However, the ignorance of the rival classes makes the training of our classifier very efficient. As mentioned in the third point, there is a big difference in the size of the optimization problem. Let the number of classes be c and the number of training samples per class be n . Then, the coefficient matrix of the quadratic optimization problem of SVM is an $nc \times nc$ dense matrix. It takes large storage and solving this optimization problem takes a large amount of computation (Osuna et al., 1997). On the other hand, since the coefficient matrix G of our classifier is an $n \times n$ matrix, the computational cost for training is much smaller. Therefore, our classifier is considered to be suitable for large scale problems such as Kanji recognition (Kimura et al., 1987).

6. Conclusion

In this paper, the subspace classifier is applied to the Hilbert space, and obtained the improvement over the conventional subspace classifier.

This improvement is due to the effect that the dimensionality of the intersection of subspaces becomes zero by the mapping. We also compared our classifier with SVM. The error rate of our classifier was not better than that of SVM, but our advantage is the small training cost.

Studies on the classification in the Hilbert space are started by SVM. But, various classifiers can be transferred to the Hilbert space, which could be very fruitful. In this paper, we have transferred the subspace classifier into the Hilbert space, and obtained a classifier with high performance and easy training. In future works, we continue to develop the subspace classifier in the Hilbert space to more sophisticated forms such as learning subspace classifiers. Also, we will apply this classifier to actual large scale problems.

References

- Bellman, R., 1970. *Introduction to Matrix Analysis*, 2nd Ed. McGraw-Hill, New York.
- Deshuang, H., 1996. The statistical properties of the learning subspace methods for pattern recognition. *Trans. Inform. Process. Soc. Jpn* 37 (6), 1081–1087.
- Haykin, S., 1995. *Neural Networks: A Comprehensive Foundation*. IEEE Computer Society Press, Silver Spring, MD.
- Kimura, F., Takashina, K., Tsuruoka, S., Miyake, Y., 1987. Modified quadratic discriminant functions and the application to chinese character recognition. *IEEE Trans. Patt. Anal. Mach. Intell.* 9 (1), 149–153.
- Oja, E., 1983. *Subspace Methods of Pattern Recognition*. Research Studies Press.
- Osuna, E.E., Freund, R., Girosi, F., 1997. Support vector machines: Training and applications. MIT AI Memo 1602.
- Prakash, M., Murty, M.N., 1996. Extended subspace methods of pattern recognition. *Pattern Recognition Letters* 17 (11), 1131–1139.
- Prakesh, M., Murty, M.N., 1997. Hebbian learning subspace method: A new approach. *Pattern Recognition* 30 (1), 141–149.
- Small, C.G., 1994. *Hilbert Space Methods in Probability and Statistical Inference*. Wiley, New York.
- Scholkopf, B., Sung, K.K., Burges, C.J.C., Girosi, F., Niyogi, P., Poggio, T., Vapnik, V., 1997. Comparing support vector machines with gaussian kernels to radial basis function classifiers. *IEEE Trans. Signal Process.* 45 (11), 2758–2765.
- Vapnik, V.N., 1995. *The Nature of Statistical Learning Theory*. Springer, Berlin.