A Support Vector Machine Formulation to PCA Analysis and Its Kernel Version

J. A. K. Suykens, T. Van Gestel, J. Vandewalle, and B. De Moor

***Abstract*—In this letter, we present a simple and straightforward primal-dual support vector machine formulation to the problem of principal component analysis (PCA) in dual variables. By considering a mapping to a high-dimensional feature space and application of the kernel trick (Mercer theorem) kernel PCA is obtained as introduced by Schölkopf *et al.* While least squares support vector machine classifiers have a natural link with kernel Fisher discriminant analysis (minimizing the within class scatter around targets $+1$ and -1), for PCA analysis one can take the interpretation of a one-class modeling problem with zero target value around which one maximizes the variance. The score variables are interpreted as error variables within the problem formulation. In this way primal-dual constrained optimization problem interpretations to linear and kernel PCA analysis are obtained in a similar style as for least square-support vector machine (LS-SVM) classifiers.**

***Index Terms*—Kernel methods, kernel principal component analysis (PCA), least squares-support vector machine (LS-SVM), PCA analysis, SVMs.**

I. Introduction

Support vector machines (SVMs) as originally introduced by Vapnik within the area of statistical learning theory and structural risk minimization [23] have proven to work successfully on many applications of nonlinear classification and function estimation. The problems are formulated as convex optimization problems, usually quadratic programs, for which the dual problem is solved. Within the models and the formulation one makes use of the kernel trick which is based on the Mercer theorem related to positive definite kernels. One can plug in any positive definite kernel for a support vector machine classifier or regressor with as typical choices linear, polynomial and RBF kernels. The work on SVMs has also stimulated the research on kernel-based learning methods in general in recent years [18]. The conceptual idea of generalizing an existing linear technique to a nonlinear version by applying the kernel trick has become an area of active research. One important result in this direction is the extension of linear principal component analysis (PCA) [11] to kernel PCA, as shown by Schölkopf *et al.* [16], [17].

The aim of this paper is to present a new simple and straightforward formulation to PCA analysis and its kernel version. The formulation is in the style of SVMs, in the sense that one starts from a constrained optimization problem in primal weight space with incorporation of a regularization term and one solves the dual problem after application of the kernel trick. The nonlinear version of the formulation yields a solution which is equivalent to kernel PCA.

The formulation is made in a similar fashion as least squares support vector machine (LS-SVM) classifiers [19]. For classification there is a close connection between LS-SVMs and kernel Fisher discriminant analysis [1], [12], [18], [22] as the within class scatter is minimized around targets $+1$ and -1. The PCA analysis problem is interpreted as a one-class modeling problem with a target value equal to zero around which one maximizes the variance. This results into a sum squared error cost function with regularization. The score variables are taken as additional error variables. As a result, this paper shows an extension of LS-SVM formulations to the area of unsupervised learning. The LS-SVM approach is closely related to regularization networks, Gaussian processes, kernel ridge regression and reproducing kernel Hilbert spaces (RKHS) [4], [14], [15], [24], [25]. On the other hand, the LS-SVM formulations are closer related to standard SVMs with explicit primal-dual interpretations from the viewpoint of optimization theory. Extensions of LS-SVMs have been given also to recurrent networks and control [21]. Issues of robustness and sparseness have been discussed in [20]. For the support vector machine interpretation to PCA analysis in this paper, sparseness can be obtained by considering it in relation to a reduced set approach [17] or a Nyström approximation [26].

This paper is organized as follows. In Section II we briefly state the classical and well-known problem of linear PCA analysis. In Section III we present the new support vector machine formulation to linear PCA in dual variables. Finally, in Section IV the nonlinear version is given which leads to kernel PCA.

II. Classical PCA Formulation

Consider a given set of data $\{x_k\}_{k=1}^N$ with $x_k \in \mathbb{R}^n$ and N given data points for which one aims at finding projected variables $w^T x_k$ with maximal variance [9], [10], [11], [13]. This means

$$\max_w \operatorname{Var}(w^T x) = \operatorname{Cov}(w^T x,\, w^T x) \simeq w^T C w \tag{1}$$

where $C = (1/N)\sum_{k=1}^N x_k x_k^T$. One optimizes this objective function under the constraint that $w^T w = 1$. This gives the Lagrangian $\mathcal{L}(w;\ \lambda) = (1/2) w^T C w - \lambda(w^T w - 1)$ with Lagrange multiplier λ. The solution follows then from $\partial\mathcal{L}/\partial w = 0$, $\partial\mathcal{L}/\partial\lambda = 0$ and is given by the eigenvalue problem

$$Cw = \lambda w. \tag{2}$$

The matrix C is symmetric and positive semidefinite. The eigenvector w corresponding to the largest eigenvalue determines the projected variable with maximal variance. Efficient and reliable numerical methods are discussed, e.g., in [7]. Neural network approaches to PCA analysis are discussed e.g., in [3], [5].

III. SVM Formulation to Linear PCA

Let us now reformulate the PCA problem as follows:

$$\max_w \sum_{k=1}^N (0 - w^T x_k)^2 \tag{3}$$

where zero is considered as a single target value. While for Fisher discriminant analysis one considers two target values $+1$ and -1 that represent the two classes, in the PCA analysis case a zero target value

Manuscript received April 16, 2002. This work was supported by the Research Council Katholieke Universiteit Leuven Concerted Research Action GOA-Mefisto 666 (Mathematical Engineering), IDO (IOTA Oncology, Genetic networks), the Flemish Government Fund for Scientific Research Flanders under Projects G.0256.97 (subspace), G.0115.01 (bio-i and microarrays), G.0240.99 (multilinear algebra), G.0197.02 (power islands), and G.0407.02 (support vector machines), research communities ICCoS, ANMMM), AWI (Bil. Int. Collaboration Hungary/Poland), IWT (Soft4s (softsensors), STWW-Genprom (gene promotor prediction), GBOU McKnow (Knowledge management algorithms), Eureka-Impact (MPC-control), Eureka-FLiTE (flutter modeling), the Belgian Federal Government under Grants DWTC IUAP IV-02 (1996-2001) and IUAP V-10-29 (2002-2006), the Dynamical Systems and Control: Computation, Identification, and Modeling under Program Sustainable Development PODO-II (CP-TR-18 Sustainibility effects of Traffic Management Systems); Direct contract research: Verhaert, Electrabel, Elia, Data4s, IPCOS. This research work was carried out at the ESAT laboratory and the Interdisciplinary Center of Neural Networks ICNN of the K. U. Leuven.

The authors are with the Katholieke Universiteit Leuven, Department of Electrical Engineering, ESAT-SCD-SISTA, B-3001 Leuven (Heverlee), Belgium (e-mail: johan.suykens@esat.kuleuven.ac.be).

Digital Object Identifier 10.1109/TNN.2003.809414

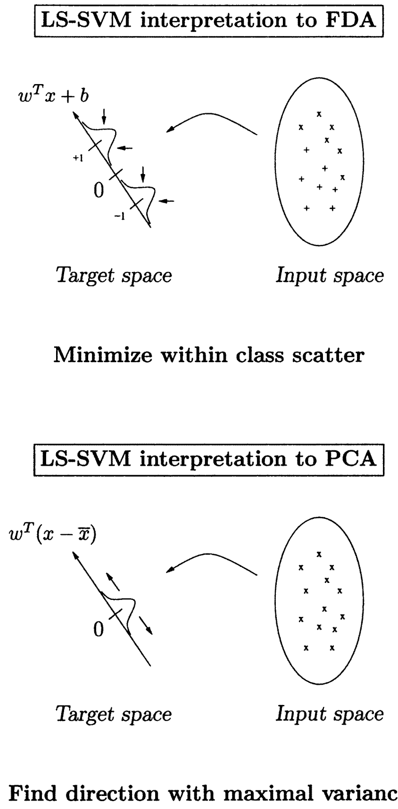
LS-SVM interpretation to FDA
$w^T x + b$
+1
0
-1
Target space
Input space
Minimize within class scatter
LS-SVM interpretation to PCA
$w^T(x - \overline{x})$
0
Target space
Input space
Find direction with maximal variance

Fig. 1. Both Fisher discriminant analysis (FDA) (supervised learning) and PCA analysis (unsupervised learning) can be derived from the viewpoint of LS-SVMs as a constrained optimization problem formulated in the primal space and solved in the dual space of Lagrange multipliers. In FDA the within class scatter is minimized around targets +1 and -1. PCA analysis can be interpreted as maximizing the variance around target 0, i.e., as a one-class target zero modeling problem.

is considered. Hence, one has in fact a one class modeling problem, but with a different objective function in mind. For Fisher discriminant analysis one aims at minimizing the within scatter around the targets, while for PCA analysis one is interested in finding the direction(s) for which the variance is maximal (Fig. 1).

This interpretation of the problem leads to the following primal optimization problem

$$\boxed{\mathrm{P}} : \max_{w,\, e} J_{\mathrm{P}}(w,\, e) = \gamma \tfrac{1}{2} \sum_{k=1}^{N} e_k^2 - \tfrac{1}{2} w^T w$$

$$\text{such that } e_k = w^T x_k, \quad k = 1, \ldots, N. \tag{4}$$

This formulation states that one considers the difference between $w^T x_k$ (the projected data points to the target space) and the value 0 as error variables. The projected variables correspond to what one calls the *score* variables. These error variables are maximized for the given N data points while keeping the norm of w small by the regularization term. The value γ is a positive real constant. The Lagrangian becomes $\mathcal{L}(w,\, e;\, \alpha) = \gamma(1/2)\sum_{k=1}^{N} e_k^2 - (1/2)w^T w - \sum_{k=1}^{N} \alpha_k(e_k - w^T x_k)$ with conditions for optimality given by

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial w} = 0 \rightarrow w = \displaystyle\sum_{k=1}^{N} \alpha_k x_k & \\ \dfrac{\partial \mathcal{L}}{\partial e_k} = 0 \rightarrow \alpha_k = \gamma e_k, & k = 1, \ldots, N \\ \dfrac{\partial \mathcal{L}}{\partial \alpha_k} = 0 \rightarrow e_k - w^T x_k = 0, & k = 1, \ldots, N. \end{cases} \tag{5}$$

By elimination of the variables $e,\, w$ one obtains $(1/\gamma)\alpha_k - \sum_{l=1}^{N} \alpha_l x_l^T x_k = 0$ for $k = 1, \ldots, N$. By defining $\lambda = 1/\gamma$ one has the following dual symmetric eigenvalue problem

$$\boxed{\mathrm{D}} : \text{ solve in } \alpha:$$

$$\begin{bmatrix} x_1^T x_1 & \cdots & x_1^T x_N \\ \vdots & & \vdots \\ x_N^T x_1 & \cdots & x_N^T x_N \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{bmatrix} = \lambda \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{bmatrix} \tag{6}$$

which is the dual interpretation of (2). The vector of dual variables $\alpha = [\alpha_1;\, \ldots;\, \alpha_N]$ is an eigenvector of the Gram matrix and λ is the corresponding eigenvalue. In order to obtain the maximal variance one selects the eigenvector corresponding to the largest eigenvalue.

The score variables become

$$z(x) = w^T x = \sum_{l=1}^{N} \alpha_l x_l^T x \tag{7}$$

where α is the eigenvector corresponding to the largest eigenvalue for the first score variable.

Remarks:

- Note that all eigenvalues are positive and real because the matrix is symmetric and positive definite. One has in fact N local minima as solution to the problem for which one selects the solution of interest. The optimal solution is the eigenvector corresponding to the largest eigenvalue because in that case $\sum_{k=1}^{N} (w^T x_k)^2 = \sum_{k=1}^{N} e_k^2 = \sum_{k=1}^{N} (1/\gamma^2)\alpha_k^2 = \lambda_{\max}^2$, where $\sum_{k=1}^{N} \alpha_k^2 = 1$ for the normalized eigenvector. For the different score variables one selects the eigenvectors corresponding to the different eigenvalues. The score variables are decorrelated from each other due to the fact that the α eigenvectors are orthonormal. According to [11], one can also additionally stress within the constraints of the formulation that the w vectors related to subsequent scores are orthogonal to each other.
- The normalization $\sum_{k=1}^{N} \alpha_k^2 = 1$ leads to $w^T w = \lambda$ in the primal space which is different from the constraint $w^T w = 1$ in (1)–(2).
- PCA analysis is usually applied to centered data. Therefore one better considers the problem $\max_w \sum_{k=1}^{N} [w^T(x_k - \hat{\mu}_x)]^2$ where $\hat{\mu}_x = (1/N)\sum_{k=1}^{N} x_k$. The same derivations can be made and one finally obtains a centered Gram matrix as a result. One also sees that solving the problem in w is typically advantageous for large data sets, while for fewer given data in huge dimensional input spaces one better solves the dual problem. The approach of taking the eigenvalue decomposition of the centered Gram matrix is also done in principal coordinate analysis [8], [11].
- In these formulations, it is also straightforward to work with a bias term in the formulation and takes $z(x) = w^T x + b$ and the primal problem $\max_{w,\, e} J_{\mathrm{P}}(w,\, e) = \gamma(1/2)\sum_{k=1}^{N} e_k^2 - (1/2)w^T w$ such that $e_k = w^T x_k + b$, for $k = 1, \ldots, N$. In a similar way, one additionally obtains then from the conditions for optimality that $\sum_{k=1}^{N} \alpha_k = 0$ which automatically leads to an expression for the bias term $b = -(1/N)\sum_{k=1}^{N} \sum_{l=1}^{N} \alpha_l x_l^T x_k$ that corresponds to a centered Gram matrix.

IV. An LS-SVM Approach to Kernel PCA

We now follow the usual SVM methodology of mapping the data from the input space to a high-dimensional feature space and applying the kernel trick.

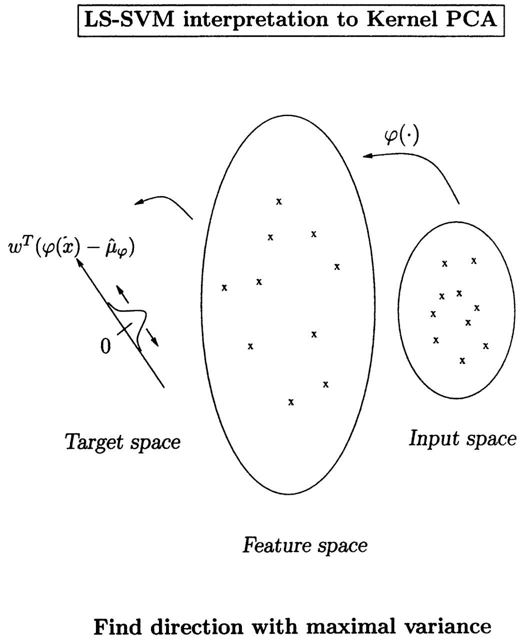
LS-SVM interpretation to Kernel PCA
$\varphi(\cdot)$
$w^T(\varphi(x) - \hat{\mu}_\varphi)$
0
Target space
Input space
Feature space
Find direction with maximal variance

Fig. 2. LS-SVM approach to kernel Fisher discriminant analysis: the input data are mapped to a high-dimensional feature space and next to the score variables. The score variables are interpreted as error variables in a one-class modeling problem with target zero for which one aims at having maximal variance.

Our objective is the following:

$$\max_w \sum_{k=1}^{N} [0 - w^T(\varphi(x_k) - \hat{\mu}_\varphi)]^2 \tag{8}$$

with notation $\hat{\mu}_\varphi = (1/N)\sum_{k=1}^{N} \varphi(x_k)$ and $\varphi(\cdot)$: $\mathbb{R}^n \rightarrow \mathbb{R}^{n_h}$ the mapping to a high-dimensional feature space which might be infinite dimensional (Fig. 2). We take here the centering approach instead of using a bias term in the formulation. The following optimization problem is formulated now in the primal weight space:

$$\boxed{\text{P}}: \max_{w,\, e} J_{\text{P}}(w,\, e) = \gamma \tfrac{1}{2} \sum_{k=1}^{N} e_k^2 - \tfrac{1}{2} w^T w$$
$$\text{such that } e_k = w^T(\varphi(x_k) - \hat{\mu}_\varphi), \quad k = 1, \ldots, N. \tag{9}$$

This gives the Lagrangian $\mathcal{L}(w,\, e;\; \alpha) = \gamma(1/2)\sum_{k=1}^{N} e_k^2 - (1/2)w^T w - \sum_{k=1}^{N} \alpha_k \left(e_k - w^T(\varphi(x_k) - \hat{\mu}_\varphi)\right)$ with conditions for optimality

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial w} = 0 \rightarrow w = \displaystyle\sum_{k=1}^{N} \alpha_k(\varphi(x_k) - \hat{\mu}_\varphi) & \\ \dfrac{\partial \mathcal{L}}{\partial e_k} = 0 \rightarrow \alpha_k = \gamma e_k, & k = 1, \ldots, N \\ \dfrac{\partial \mathcal{L}}{\partial \alpha_k} = 0 \rightarrow e_k - w^T(\varphi(x_k) - \hat{\mu}_\varphi) = 0, & k = 1, \ldots, N. \end{cases} \tag{10}$$

By elimination of the variables $e,\, w$ one obtains $(1/\gamma)\alpha_k - \sum_{l=1}^{N} \alpha_l(\varphi(x_l) - \hat{\mu}_\varphi)^T(\varphi(x_k) - \hat{\mu}_\varphi) = 0, k = 1, \ldots, N$. Defining $\lambda = 1/\gamma$ one obtains the following dual problem:

$$\boxed{\text{D}}: \text{ solve in } \alpha\text{: } \Omega_c \alpha = \lambda \alpha \tag{11}$$

with (12) shown at the bottom of the page. One has the following elements for the centered kernel matrix:

$$\Omega_{c,\, kl} = (\varphi(x_k) - \hat{\mu}_\varphi)^T(\varphi(x_l) - \hat{\mu}_\varphi), \qquad k,\, l = 1, \ldots, N. \tag{13}$$

For the centered kernel matrix one can apply the kernel trick as follows for given points $x_k,\, x_l$

$$\begin{aligned} &(\varphi(x_k) - \hat{\mu}_\varphi)^T\, (\varphi(x_l) - \hat{\mu}_\varphi) \\ &\quad = K(x_k,\, x_l) - \frac{1}{N}\sum_{r=1}^{N} K(x_k,\, x_r) \\ &\qquad - \frac{1}{N}\sum_{r=1}^{N} K(x_l,\, x_r) + \frac{1}{N^2}\sum_{r=1}^{N}\sum_{s=1}^{N} K(x_r,\, x_s) \end{aligned} \tag{14}$$

with application of the kernel trick $K(x_k,\, x_l) = \varphi(x_k)^T\varphi(x_l)$ based on the Mercer theorem. A typical choice is the RBF kernel $K(x_k,\, x_l) = \exp(-\|x_k - x_l\|_2^2/\sigma^2)$. This solution is equivalent with the kernel PCA solution as proposed by Schölkopf *et al.* in [16]. The centered kernel matrix can be computed as $\Omega_c = M_c \Omega M_c$ with $\Omega_{kl} = K(x_k,\, x_l)$ and centering matrix $M_c = I - (1/N)1_v 1_v^T$ with $1_v = [1;\; 1;\; \ldots;\; 1]$. This issue of centering is also of importance in methods of principal coordinate analysis [11].

The optimal solution to the formulated problem is obtained by selecting the eigenvector corresponding to the largest eigenvalue. The projected variables become

$$\begin{aligned} z(x) &= w^T\, (\varphi(x) - \hat{\mu}_\varphi) \\ &= \sum_{l=1}^{N} \alpha_l \left(K(x_l,\, x) - \frac{1}{N}\sum_{r=1}^{N} K(x_r,\, x) \right. \\ &\qquad - \frac{1}{N}\sum_{r=1}^{N} K(x_r,\, x_l) \\ &\qquad \left. + \frac{1}{N^2}\sum_{r=1}^{N}\sum_{s=1}^{N} K(x_r,\, x_s) \right). \end{aligned} \tag{15}$$

Similar remarks hold as given for the linear case. For the nonlinear PCA case the selected number of score variables n_s can be larger than the dimension of the input space n. One selects then as few score variables as possible and minimizes the reconstruction error [17]. Furthermore, the link between kernel PCA and density estimation has been recently discussed in [6]. Sparseness in this support vector machine formulation can be related to considering a reduced set approach [17], also in relation to the Nyström method [26]. The advantage of having primal-dual interpretations is that the dual problem is suitable for handling large-dimensional input spaces, while the primal problem is better to treat problems with a large number of data N.

$$\Omega_c = \begin{bmatrix} (\varphi(x_1) - \hat{\mu}_\varphi)^T(\varphi(x_1) - \hat{\mu}_\varphi) & \cdots & (\varphi(x_1) - \hat{\mu}_\varphi)^T(\varphi(x_N) - \hat{\mu}_\varphi) \\ \vdots & & \vdots \\ (\varphi(x_N) - \hat{\mu}_\varphi)^T(\varphi(x_1) - \hat{\mu}_\varphi) & \cdots & (\varphi(x_N) - \hat{\mu}_\varphi)^T(\varphi(x_N) - \hat{\mu}_\varphi) \end{bmatrix}. \tag{12}$$

V. Conclusion

A new SVM style formulation has been given to PCA analysis. The use of a mapping to a high-dimensional feature space leads to the kernel PCA version of Schölkopf *et al.* Conceptually, the formulation considers the problem as a one-class modeling problem with zero target value around which one maximizes the variance. A straightforward comparison can be made with the problem of Fisher discriminant analysis where the within class scatter is minimized around target values $+1$ and -1. Natural links exist with LS-SVM classifiers. This result also further extends LS-SVM methods toward unsupervised learning.

References

[1] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Comput.*, vol. 12, pp. 2385–2404, 2000.

[2] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Oxford Univ. Press, 1995.

[3] K. I. Diamantaras and S. Y. Kung, *Principal Component Neural Networks*. New York: Wiley, 1996.

[4] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization networks and support vector machines," *Adv. Comput. Math.*, vol. 13, no. 1, pp. 1–50, 2000.

[5] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. San Diego, CA: Academic, 1990.

[6] M. Girolami, "Orthogonal series density estimation and the kernel eigenvalue problem," *Neural Comput.*, vol. 14, no. 3, pp. 669–688, 2002.

[7] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore, MD: Johns Hopkins Univ. Press, 1989.

[8] J. C. Gower, "Some distance properties of latent root and vector methods used in multivariate analysis," *Biometrika*, vol. 53, pp. 325–338, 1966.

[9] H. Hotelling, "Simplified calculation of principal components," *Psychometrica*, vol. 1, pp. 27–35, 1936.

[10] ——, "Relations between two sets of variates," *Biometrica*, vol. 28, pp. 321–377, 1936.

[11] I. T. Jolliffe, *Principal Component Analysis*, ser. Springer Series in Statistics. New York: Springer-Verlag, 1986.

[12] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller, "Fisher discriminant analysis with kernels," in *Neural Networks for Signal Processing IX*, Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, Eds. Piscataway, NJ: IEEE Press, 1999, pp. 41–48.

[13] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Phil. Mag.*, vol. 6, no. 2, pp. 559–572, 1901.

[14] T. Poggio and F. Girosi, "Networks for approximation and learning," *Proc. IEEE*, vol. 78, no. 9, pp. 1481–1497, 1990.

[15] C. Saunders, A. Gammerman, and V. Vovk, "Ridge regression learning algorithm in dual variables," in *Proc. 15th Int. Conf. Machine Learning ICML-98*, Madison, WI, 1998, pp. 515–521.

[16] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, pp. 1299–1319, 1998.

[17] B. Schölkopf, S. Mika, C. Burges, P. Knirsch, K.-R. Müller, G. Rätsch, and A. Smola, "Input space vs. feature space in kernel-based methods," *IEEE Trans. Neural Networks*, vol. 10, pp. 1000–1017, Sept. 1999.

[18] B. Schölkopf and A. Smola, *Learning With Kernels*. Cambridge, MA: MIT Press, 2002.

[19] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Lett.*, vol. 9, no. 3, pp. 293–300, 1999.

[20] J. A. K. Suykens, J. De Brabanter, L. Lukas, and J. Vandewalle, "Weighted least squares support vector machines: Robustness and sparse approximation," *Neurocomputing (Special Issue on Fundamental and Information Processing Aspects of Neurocomputing)*, vol. 48, no. 1–4, pp. 85–105, 2002.

[21] J. A. K. Suykens, J. Vandewalle, and B. De Moor, "Optimal control by least squares support vector machines," *Neural Networks*, vol. 14, no. 1, pp. 23–35, 2001.

[22] T. Van Gestel, A. J. K. Suykens, G. Lanckriet, A. Lambrechts, B. De Moor, and J. Vandewalle, "A Bayesian framework for least squares support vector machine classifiers, Gaussian processes and kernel Fisher discriminant analysis," *Neural Comput.*, vol. 14, no. 5, pp. 1115–1147, 2002.

[23] V. Vapnik, *The Nature of Statistical Learning Theory*. New-York: Springer-Verlag, 1995.

[24] G. Wahba, *Spline Models for Observational Data*, ser. Series in Applied Mathematics. Philadelphia, PA: SIAM, 1990, vol. 59.

[25] C. K. I. Williams and C. E. Rasmussen, "Gaussian processes for regression," in *Advances in Neural Information Processing Systems*, D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, Eds. Cambridge, MA: MIT Press, 1996, vol. 8, pp. 514–520.

[26] C. K. I. Williams and M. Seeger, "Using the Nyström method to speed up kernel machines," in *Advances in Neural Information Processing Systems*, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds. Cambridge, MA: MIT Press, 2001, vol. 13, pp. 682–688.

On the Capability of Accommodating New Classes Within Probabilistic Neural Networks

Tetsuya Hoya

***Abstract*—To date, probabilistic neural networks (PNNs) have been widely used in various pattern classification tasks due to their robustness. In this letter, it is shown that by exploiting the flexible network configuration property, the PNN classifiers also exhibit the capability in accommodating new classes. This is verified by extensive simulation studies on using four different domain data sets for pattern classification tasks.**

***Index Terms*—Accomodating new classes, incremental learning, pattern classification, probabilistic neural networks (PNNs).**

I. Introduction

PROBILISTIC neural networks (PNNs) [1] /generalized regression neural networks (GRNNs) [2] belong to the family of radial basis function neural networks (RBF-NNs) [3], [4], which, due to their robustness, are now widely used in various pattern classification tasks. Although the roots of these two networks are strictly and statistically not the same, they exhibit similar characteristics to each other and share the attractive property, namely, that they require no iterative training of the weight vector between the RBFs and the output units [5]. By exploiting this property, the author has proposed an instance-based incremental training scheme using a GRNN [6] in which pattern correction of misclassification data was performed by an on-line batch correction mechanism.

In a recent study [7], a new guideline for the incremental learning paradigm in pattern classification has been given in accordance with the four criteria.

1) The pattern classifier(s) should be able to learn additional information from the new data.
2) They should not require access to the original data, used to train the existing classifier.
3) They should preserve previously acquired knowledge (that is, they should not suffer from catastrophic forgetting).
4) They should be able to accommodate new classes that may be introduced within the new data.

Manuscript received April 29, 2002; revised August 25, 2002 and September 29, 2002.

The author is with the Laboratory for Advanced Brain Signal Processing, Brain Science Institute, RIKEN, Wakoh-City, Saitama 351-0198, Japan (e-mail: hoya@bsp.brain.riken.go.jp).

Digital Object Identifier 10.1109/TNN.2003.809417

1045-9227/03$17.00 © 2003 IEEE