Kernel-Based Associative Memory*

Dimitri Nowicki
Institute of Mathematical
Machines and Systems of NASU
Kiev, Ukraine
E-mail: nowicki@fromru.com

Oleksiy Dekhtyarenko
Institute of Mathematical
Machines and Systems of NASU
Kiev, Ukraine
E-mail: olexii@mail.ru

Abstract– **We propose a new approach to pseudo-inverse associative memories using kernel machine methodology. Basing on Hopfield-type pseudoinverse associative memories we developed a series of kernel-based hetero- and auto-associative algorithms. There are convergence processes possible during examination procedures even for continuous data. Kernel approach enables to overcome capacity limitations inherent to Hopfield-type networks. Memory capacity virtually does not depend on data dimension. We provide theoretical investigation for proposed methods and prove its attraction properties. Also we have experimentally tested them for tasks of classification and associative retrieval.**

I. Introduction

Nowadays there is a drastic growth in the domain of kernel machines. These methods and techniques are widely applied for pattern recognition, regression, classification and clustering. All kernel methods use feature space whose dimension is significantly larger than the dimension of original space. Original and feature spaces are connected by the nonlinear mapping:

$$\varphi : E_X \to E'_X.$$

Dimensionality of E_X might be very large or even infinite. So it is difficult or impossible to operate with vectors in this space explicitly. However sometimes one needs to know only inner (scalar) product in E_X as a function of elements of E_X. This function is called kernel:

$$K(u,v) = (\varphi(u), \varphi(v)) \qquad (1)$$

Most of kernel methods like SVM [1], LS-SVM [2] are used for classification, regression, and pattern recognition. Thus it will be interesting to construct kernel machines with other functions such as associative recall. The aim of this paper is to construct associative memory (AM) based on kernel machine.

Linear independence of memorized vectors is required by Hopfield-type associative memories. Moreover, their linear independence must be "sufficiently strong" i.e. every vector must be sufficiently far from the linear hull of all remaining ones. It implies that the number of patterns to memorize must be less than their dimension. In practice this number does not exceed 25% (pseudoinverse learning rule, [3]) or 70% of vectors' dimension (desaturation technique, see [4]). Diagonal elements of the synaptic matrix dominate if the number of patterns is close to this limit. This implies drastic decrease of attraction properties and deterioration of associative memory's capabilities.

Using kernel machines one can change over to the space where memorized data set becomes linearly independent. We use pseudoinverse hetero- and auto-associative memory as a prototype. Usage of kernel machines in scope of this paradigm enables to overcome limitations due to linearity of the basic model. In particular, we can remove capacity limitations of these memories. Using this approach we also constructed associative memory capable to iterative convergence during examination process with the continuous data.

II. The Algorithm

Lets consider pseudoinverse heteroassociative memory. Suppose E_X and E_X are input and output spaces with dimensionalities n and p respectively. We should store m pairs of vectors $x_i \in E_X, y_i \in E_Y, i = \overline{1,m}$. These vectors are supposed to form columns of matrices $\mathbf{X}$ and $\mathbf{Y}$ respectively. In order to provide appropriate heteroassociative behavior matrix B can be specified as:

$$\mathbf{BX} = \mathbf{Y}$$

whose solution is:

$$\mathbf{B} = \mathbf{Y}\mathbf{X}^{+} \qquad (2)$$

This matrix defines a projective operator $\mathbf{B} : E_X \to E_Y$ such that $\mathbf{B}\mathbf{x}_i = \mathbf{y}_i$ for all i. We denote by operator "$^+$" the Moore-Penrose pseudoinverse of $\mathbf{X}$ (see e.g. [5]). In case of linearly independent columns pseudoinverse matrix can be found as

$$\mathbf{X}^{+} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{S}^{-1}\mathbf{X}^T. \qquad (3)$$

The elements of $m \times m$-sized matrix S are computed as pairwise scalar products of memorized vectors:

$$s_{ij} = (x_i, x_j). \qquad (4)$$

*This research was supported by INTAS grant #01-0257

Examination procedure takes an arbitrary input vector **x**. We should produce network's response **y**. This cold be done as follows:

$$\begin{aligned} y &= \mathbf{B}x = \mathbf{Y}\mathbf{S}^{-1}\mathbf{z}; \\ \mathbf{z} &= \mathbf{X}^T\mathbf{x}; \\ z_i &= (\mathbf{x}_i, \mathbf{x}). \end{aligned} \tag{5}$$

Note that we need to know only scalar products of memorized vectors themselves and input vector **x**.

We use this property of heteroassociative memory to construct the kernel algorithm. We replace E_X by E'_X whose dimensionality is $n' >> n$ (E'_X may also be an infinite-dimensional Hilbert space). Vectors in E'_X are evaluated using nonlinear transformation $\varphi: E_X \to E'_X$. E'_X is called *feature space.*

Let $x_i' = \varphi(x_i), x_i \in E_X$ be input vectors of training dataset, and $K(\mathbf{u}, \mathbf{v}) = (\varphi(\mathbf{u}), \varphi(\mathbf{v}))$ be a kernel.

Then, like (3-5) we get:

$$\begin{aligned} s_{ij} &= K(x_i, x_j); \\ z_i &= K(x_i, x). \end{aligned} \tag{6}$$

Expressions (2-3, 5-6) could be evaluated by means of kernel only, without explicit usage of E'_X, this leading to kernel-based procedures of learning and examination.

This is a basic algorithm for kernel associative memory.

Corollary 1 of Mercer's theorem: If $K(u,v)$ is a Mercer's kernel [1] then

1) Hilbert space E'_X and a mapping $\varphi: E_X \to E'_X$ exist such that $K(u, v) = (\varphi(u), \varphi(v))$
2) For each set of pairs $x_i \in E_X, y_i \in E_Y, i = 1\text{K}\, m$ matrix **S** is nonnegative-defined
3) If in addition $\dim(E_X') > m$, there exists a operator $B: E'_X \to E_Y$ such that $Bx'_i = y_i$ for all i.

Proof:

1) follows directly from Mercer's theorem
2) this is true because the matrix **S** consists of pairwise scalar products of $x_i' \in E'_X$ (it is a Gram matrix of this set of vectors)
3) such an operator could be built on the (m-dimensional) linear hull of $< \mathbf{x}_i' >_{i=1}^m \in E'_X$ and extended continuously to the whole E'_X.□

Mercer's condition is formulated as follows. Let $K(\mathbf{u}, \mathbf{v}): Q \times Q \to \Re$ be a continuous symmetric function and Q be a compact set in E_X. Then, a space E'_X and a mapping $\varphi: E_X \to E'_X$ such that

$\forall u, v \in Q \subset E_X \quad K(\mathbf{u}, \mathbf{v}) = \langle \varphi(\mathbf{u}), \varphi(\mathbf{v}) \rangle_{E'_X}$ exist if and only if for any $g \in L_2(Q)$ following inequality holds:

$$\iint_{u,v \in Q} K(u,v) g(u) g(v) \geq 0$$

Unfortunately, we cannot guarantee non-singularity of the matrix **S**. It is invertible if $< x_i' >_{i=1}^m \in E'_X$ are linearly independent. This condition may not hold for certain kernels and specific vector sets. In practice, one can suppress this problem using Tikhonov's regularization: instead of **S** using the matrix:

$$\mathbf{S}_\mu = \mathbf{S} + \mu \mathbf{I}$$

for small μ>0. This matrix is always invertible since **S** is nonnegative definite.

Another approach to this problem uses incremental construction of the matrix **S**. During each step of the algorithm its dimensionality is increased by one with the addition of each next memorized vector. If this leads to singular matrix, the vector is rejected. For inversion of **S** we use the technique for block matrices [5,6].

To memorize m patterns in this network we need to store $m \times m$-sized matrix **S**. We can say that kernel associative memory is capable to store as many images as neurons it has. This is a maximum estimation which is sometimes unreachable in practice. For instance, in case of *scalar-product* kernel this machine is identical to conventional neural heteroassociative memory.

III. Modifications of the Kernel Algorithm

A. Autoassociative memory

The algorithm described above might be also used for autoassociative memory. In this case E_x and E_y are identical, $x_i \in E_X, y_i \in E_Y, x_i = y_i, i = 1\text{K}\, m$. Matrix **S** is calculated by formula (6). There is an iterative examination procedure: the vector x_t is sent to the network's input, using (5-6) we obtain postsynaptic potential y_t. Then, in case of bipolar data we apply activation function and compute the next state of the system:

$$x_{t+1} = f(y_t) \tag{7}$$

This procedure is iterated until a stable state (attractor) has been reached. Attractors of such systems are described by

Theorem 1. *Suppose for autoassociative memory (4-8) conditions of the corollary 1 hold, and matrix* **S** *is invertible. Then attractors of corresponding examination procedure are only fixed points or 2-cycles*

Proof: We construct energy function in the way similar to the corresponding proof for Hopfield networks:

$$E_t = -\frac{1}{2}K(x_t, y_t) \tag{8}$$

By corollary from Mercer's theorem a self-conjugated operator $C: E'_x \to E_y$ exists such that $Cx'_t=y_t$. Applying properties of scalar product in E'_x we get:

$$\begin{aligned} E_t - E_{t+1} &= -\frac{1}{2}(x'_t, Cx'_{t-1}) + \frac{1}{2}(x'_{t+1}, Cx'_t) = \\ -\frac{1}{2}(x'_{t-1}, Cx'_t) &+ \frac{1}{2}(x'_{t+1}, Cx'_t) = \\ &= \frac{1}{2}K(y_t, x_{t+1}) - \frac{1}{2}K(y_t, x_{t-1}) \end{aligned} \tag{9}$$

Since the kernel is monotonic function with respect to distance between x and y expression (9) is non-negative, it is zero if and only if the fixed point is reached.□

The scheme of examination algorithm for auto-associative memory is displayed in the fig. 1.

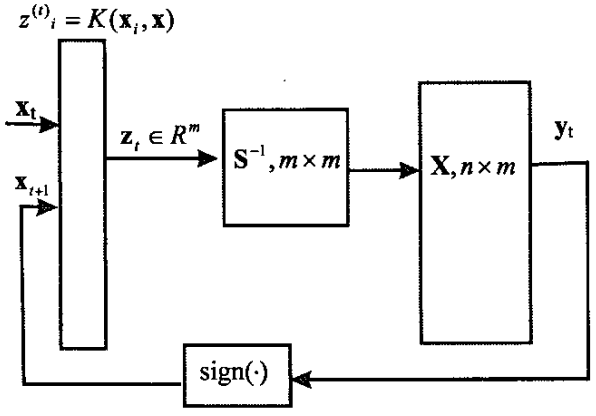

Fig. 1. Scheme of the kernel autoassociative memory

B. Internal activation function

Consider the vector $\mathbf{w} = \mathbf{S}^{-1}\mathbf{z}$. It corresponds exactly to k-th memorized pattern if and only if $w_i = \delta_{ik}$. To provide a better convergence to such w we apply *internal activation function*:

$$F: \mathbf{w} \to \mathbf{w}' \; w' = \theta(w_i)$$

where $\theta: [0,1] \to [0,1]$ is smooth monotonic function such that $\theta(0) = 0, \theta(1) = 1; \theta'(0) = \theta'(1) = 0$.

IV. Experimental Results

Our models and algorithms were experimentally tested for several tasks of auto- and heteroassociative recall and classification. Here we display results or auto-associative memory working with simulated data arrays and real-world data (images). For experiments we used following three types of kernel:

1. Polynomial
$$K(x,y) = (1 + \alpha(x,y))^{\beta}, \quad \alpha > 0, \beta - \text{positive integer} \tag{10}$$
2. Gaussian RBF
$$K(x,y) = \exp\left(-\alpha\|x-y\|^2\right), \alpha > 0 \tag{11}$$
3. Power RBF:
$$K(x,y) = \left(1 + \alpha\|x-y\|^2\right)^{\beta}, \quad \alpha > 0, \beta > 0 \tag{12}$$

We studied attraction properties of kernel associative memory. All experiments were performed using internal activation function and iterative examination procedure. Algorithms were implemented using neural-network software package NeuroLand [7].

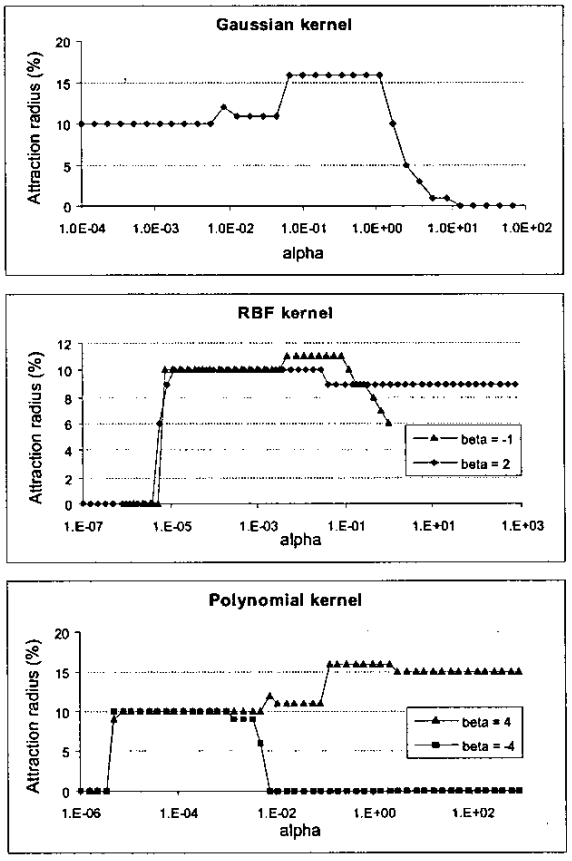

Fig. 2. Attraction radius of kernel AM for bipolar data

A. Simulated bipolar data

To study attraction properties of kernel associative memory for bipolar data we choose a network memorizing 264 64-dimensional patterns. Bipolar data

vectors were randomly generated, probabilities of values +1 and -1 for each component were equal, and components were independent.

Attraction radius was measured as a maximum value of bipolar noise such that the network still gave correct responses for all memorized patterns. In fig. 2 we display attraction radius depending on parameters for kernels (10-12).

B. Image data

In these experiments we used 30×30 gray-scale photographs of faces. They were presented as real-valued vectors with components normalized to [-1;1], the kernel AM memorized 61 images.

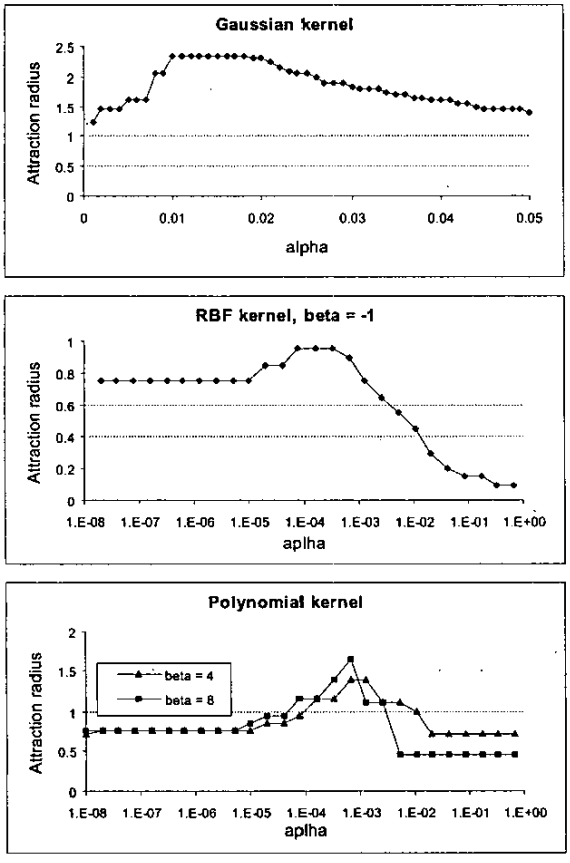

Fig. 3. Attraction radius of kernel AM for picture data

Input vectors were obtained from original patterns by adding Gaussian noise with zero mean. Standard deviation σ of this noise served as a measure of attraction radius. More precisely, attraction radius was set equal to σ such that all images were restored with fixed precision ε. Attraction radius depending on parameters for kernels (10-12) is displayed in fig. 3.

V. Conclusion

This article introduces associative memory based on kernel machine. We present theoretical justification and experimental tests for these techniques.

Unlike [8], where author uses high order generalization of the Hopfield model that includes interactions between more than two neurons, we restrict ourselves to two component Hamiltonian (energy function). Doing so we are able to provide analytical solution for the stability equation (2).

Experimental results show that proposed kernel algorithm successfully works as auto- and heteroassociative memory. We demonstrate attraction properties of kernel AM for different types of data. It may also be used for classification and pattern recognition. Using kernel methods we can construct iterative examination procedure even for continuous data. Also we can increase capacity of associative memory and overcome limitations inherent to Hopfield-type neural networks.

References

[1] V. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, NY, 1998.

[2] Smale S. On the Mathematical Foundations of Learning *Bull. Am. Math. Soc.*, Vol. 39, No. 1, pp. 1-49, 2001.

[3] L. Personnaz, I. Guyon, G. Dreyfus, "Collective computational properties of neural networks: New learning mechanisms," *Phys. Rev. A*, Vol.34 (5), pp. 4217-4228, 1986.

[4] D.O. Gorodnichy, A.M. Reznik, "Increasing Attraction of Pseudo-Inverse Neural Networks," *Neural Processing Letters,* vol. 5, pp. 121-125, 1997.

[5] A. Albert, *Regression and the Moore-Penrose pseudoinverse,* Academic Press, New York-London, 1972.

[6] L.A. Pipes, *Applied mathematics for engineers and physicists*, McGraw-Hill Book Co., New York-Toronto-London, 1958.

[7] A.M. Reznik, E.A. Kalina, A.S. Sitchov, E.G. Sadovaya, O.K. Dekhtyarenko, A.A. Galinskaya, "The multifunctional neural computer NeuroLand," *Proceedings of the Int. Conf. on Inductive Simulation,* Lviv, Ukraine, vol.1 (4), pp. 82-88. May 20-25, 2002.

[8] Barbara Caputo, "Storage Capacity of Kernel Associative Memories," *Proceedings of the Int. Conf. on Art. Neural Networks,* Aug. 27-31 2002, Madrid, Spain.