A Note on the Regularization Algorithm*

Wojciech Jaworski

Faculty of Mathematics, Computer Science and Mechanics
Warsaw University, Banacha 2, 02-07 Warsaw, Poland
wj189182@students.mimuw.edu.pl

Abstract. Regularization Algorithm (also called Regularization Network) is a technique for solving problems of learning from examples – in particular, the problem of approximating a multivariate function from sparse data. We analyze behavior of Regularization Algorithm for regularizator parameter equal to zero. We propose an approximative version of algorithm in order to overcome the computational cost for large data sets. We give proof of convergence and estimation for error of approximation.

Keywords: computational learning theory, regularization algorithm, approximate regularization algorithm

1 Introduction

The purpose of this paper is to analyze properties of Regularization Algorithm (RA, for short) and present its approximate version, which is more useful for data analysis than standard version because it has the lower computational complexity.

This paper is based on concepts introduced by Smale and Poggio [1],[2],[4],[8].

RA is used to solve problems of learning from examples, in particular problems of approximating a multivariate function from sparse data. The algorithm is searching for a function that minimizes a functional

$$\mathcal{E}_{\gamma,\mathbf{z}}(f) = \frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2 + \gamma\|f\|_K^2,$$

where $\mathbf{z} = ((x_1, y_1), \ldots, (x_m, y_m))$ is a sample of the length m, $x_i \in \mathbb{R}^n$ are examples and $y_i \in \mathbb{R}$ are decisions, $\|\cdot\|_K$ is the norm in the reproductive kernel Hilbert space. The function $f_{\gamma,\mathbf{z}}$ minimizing $\mathcal{E}_{\gamma,\mathbf{z}}$ is given by

$$f_{\gamma,\mathbf{z}}(x) = \sum_{i=1}^{m} a_i K(x, x_i),$$

where $\mathbf{a} = (a_1, \ldots, a_m)$ is a solution of a system of linear equations in $\mathbb{R}^m$:

$$(\gamma m I + K[\mathbf{x}])\mathbf{a} = \mathbf{y}.$$

* This paper consists of a part of my Master Thesis supervised by A. Skowron.

S. Tsumoto et al. (Eds.): RSCTC 2004, LNAI 3066, pp. 235–243, 2004.
© Springer-Verlag Berlin Heidelberg 2004

Derivation of the above formula is included in Section 2.

The main results of the paper concern of the RA behavior for parameter γ equal to zero and derivation of the approximate algorithm that not requires solving the above linear equation system.

In Section 3 we show that if examples in a sample are pairwise different and $\gamma = 0$ then $f_{\gamma,\mathbf{z}}$ exists, $\mathcal{E}_{\gamma,\mathbf{z}}(f_{\gamma,\mathbf{z}}) = 0$, and $f_{\gamma,\mathbf{z}}$ minimizes the norm on the set of functions for which $\mathcal{E}_{\gamma,\mathbf{z}}$ has value 0. Next, the transformation that allow us to extend this result to all sets of examples is given.

In Section 4 an approximate regularization algorithm is derived. It is grounded on the expansion idea of inverse operator into geometric series and getting the first N terms of that series. Iterative formula that makes it possible to execute approximate algorithm by means of N multiplications of matrix by a vector is presented. Moreover the estimation of error of such an approximation if found.

In Conclusions, we also outline some further research directions.

2 Regularization Algorithm

In the first place we'll define a slightly more general version of RA.

Theorem 2.1. *Let ρ be a probabilistic measure over $\mathbf{z}$, $\rho_i = \rho(z_i) > 0$. Let*

$$\mathcal{E}_{\rho,\gamma,\mathbf{z}}(f) = \frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2\rho_i + \gamma\|f\|_K^2,$$

and $\gamma > 0$. Then a function $f_{\rho,\gamma,\mathbf{z}}$ that minimizes $\mathcal{E}_{\rho,\gamma,\mathbf{z}}$ can be expressed by

$$f_{\rho,\gamma,\mathbf{z}}(x) = \sum_{i=1}^{m} a_i K(x, x_i),$$

where $\mathbf{a} = (a_1, \ldots, a_m)$ is the unique solution of well-posed linear system in $\mathbb{R}^m$:

$$(\gamma R^{-1} + K[\mathbf{x}])\mathbf{a} = \mathbf{y},$$

where R is a diagonal matrix with the value ρ_i on the i-th entry of diagonal.

Proof. $\gamma > 0$, so $\gamma\|f\|_K < \infty$ iff $f \in \mathcal{H}_K$. Hence may assume that $f \in \mathcal{H}_K$ and $f = \sum c_k\phi_k$. Then $\|f\|_K^2 = \sum \frac{c_k^2}{\lambda_k}$. Let

$$H(f) = \sum_{i=1}^{m}(y_i - f(x_i))^2\rho_i + \gamma\sum_k \frac{c_k^2}{\lambda_k}.$$

For every $k \geq 1$

$$\frac{\partial H(f)}{\partial c_k} = \sum_{i=1}^{m} 2(y_i - f(x_i))(-\phi_k(x_i))\rho_i + 2\gamma\frac{c_k}{\lambda_k}.$$

If f is minimum of H, then $\frac{\partial H(f)}{\partial c_k} = 0$ for all k.
So, we obtain $c_k = \frac{\lambda_k}{\gamma} \sum_{i=1}^{m} (y_i - f(x_i))\phi_k(x_i)\rho_i$.

Let $a_i = \frac{y_i - f(x_i)}{\gamma}\rho_i$. Then $c_k = \lambda_k \sum_i a_i \phi_k(x_i)$ and

$$f(x) = \sum_{k=1}^{\infty} c_k \phi_k(x) = \sum_{k=1}^{\infty} (\lambda_k \sum_{i=1}^{m} a_i \phi_k(x_i))\phi_k(x) =$$

$$= \sum_{i=1}^{m} a_i \sum_{k=1}^{\infty} \lambda_k \phi_k(x_i)\phi_k(x) = \sum_{i=1}^{m} a_i K(x_i, x).$$

In definition of a_i we replace $f(x_i)$ with $\sum_{i=1}^{m} a_i K(x_i, x)$ and we obtain:

$$a_i = \frac{y_i - \sum_{k=1}^{m} a_i K(x_k, x_i)}{\gamma}\rho_i.$$

Hence

$$R^{-1}\mathbf{a} = \frac{\mathbf{y} - K[\mathbf{x}]\mathbf{a}}{\gamma}.$$

Corollary 2.2. *Function $f_{\gamma,\mathbf{z}}$ minimizing $\mathcal{E}_{\gamma,\mathbf{z}}$ may be expressed by*

$$f_{\gamma,\mathbf{z}}(x) = \sum_{i=1}^{m} a_i K(x, x_i),$$

where $\mathbf{a} = (a_1, \ldots, a_m)$ is the solution of linear system in $\mathbb{R}^m$:

$$(\gamma m I + K[\mathbf{x}])\mathbf{a} = \mathbf{y}.$$

Proof. Proof follows immediately from the above theorem for $\rho_i = \frac{1}{m}$.

3 Regularization Algorithm for $\gamma = 0$

K is positive definite so $K[\mathbf{x}]$ is positive, if only components of $\mathbf{x}$ are pairwise different. On the other hand, when in $\mathbf{x}$ are two identical components then $K[\mathbf{x}]$ has two identical rows, so 0 is its eigenvalue.

Regularization algorithm is executable if only $\gamma m I + K[\mathbf{x}]$ is invertible. So, if the sequence of examples $\mathbf{x}$ doesn't have any repetition then the algorithm may be executed for $\gamma = 0$.

Algorithm can be extended over all sequences of examples through cumulating identical examples: Let $\mathbf{z}' = \{(x'_1, y'_1), \ldots, (x'_{m'}, y'_{m'})\}$ be a sample such that $x'_i \neq x'_j$ for $i = 1, \ldots, m'$, and ρ a probabilistic measure defined as follows: Assume that for each $x \in \mathbf{x}$ exists $x' \in \mathbf{x}'$ such that $x = x'$. Let $A_i = \{j \in 1, \ldots, m : x'_i = x_j\}$. Then $y'_i = \frac{1}{|A_i|} \sum_{j \in A_i} y_i$, $\rho_i = \frac{|A_i|}{m}$. $\mathbf{z}'$ depends on $\mathbf{z}$ in following way:

Proposition 3.1.

$$\bigvee_{f\in\mathcal{H}_K}\ \mathcal{E}_{\gamma,\mathbf{z}}(f)=\mathcal{E}_{\rho,\gamma,\mathbf{z}'}(f)+C,$$

where C is a constant independent from f and γ.

Proof.

$$\mathcal{E}_{\gamma,\mathbf{z}}(f)=\frac{1}{m}\sum_{i=1}^{m}(f(x_i)-y_i)^2+\gamma\|f\|_K^2=\sum_{i=1}^{|\mathbf{z}'|}\sum_{j\in A_i}\frac{1}{m}(f(x_j)-y_j)^2+\gamma\|f\|_K^2$$

$$\sum_{j\in A_i}\frac{1}{m}(f(x_j)-y_j)^2=\frac{1}{m}\sum_{j\in A_i}(f(x_i')-y_j)^2=$$

$$=\frac{1}{m}(|A_i|f(x_i')^2-2\sum_{j\in A_i}f(x_i')y_j+\sum_{j\in A_i}y_j^2)=$$

$$=\frac{1}{m}(|A_i|(f(x_i')-\sum_{j\in A_i}y_j)^2-|A_i|(\sum_{j\in A_i}y_j)^2+\sum_{j\in A_i}y_j^2)=\rho_i(f(x_i')-y_i')^2+c_i.$$

It follows from the above, that it is enough to consider samples without repetitions of examples.

Then our algorithm has the following properties:

Proposition 3.2.

$$\mathcal{E}_{\mathbf{z}}(f_{\rho,0,\mathbf{z}})=\mathcal{E}_{\rho,0,\mathbf{z}}(f_{\rho,0,\mathbf{z}})=0.$$

Proof. We will show that $f_{\rho,0,\mathbf{z}}(x_j)=y_j$. We have

$$f_{\rho,0,\mathbf{z}}(x_j)=\sum_{i=1}^{m}a_iK(x_j,x_i).$$

Since $\gamma=0$, we have $K[\mathbf{x}]\mathbf{a}=\mathbf{y}$. Hence $\sum_{i=1}^{m}a_iK(x_j,x_i)=y_j$.

Theorem 3.3.

$$\|f_{\rho,0,\mathbf{z}}\|_K=\inf_{f:\mathcal{E}_{\mathbf{z}}(f)=0}\|f\|_K$$

and $f_{\rho,0,\mathbf{z}}$ is the only function for which infimum is reached.

Proof. $\mathcal{E}_{\rho,\gamma,\mathbf{z}}(f)$ is continuous with respect to $f\in\mathcal{H}_K$ and $\gamma\in\mathbb{R}$. Let $F(\gamma):(0,+\infty)\to\mathcal{H}_K$, $F(\gamma)=f_{\rho,\gamma,\mathbf{z}}$. $F(\gamma)$ is a function for which $\inf_f\mathcal{E}_{\rho,\gamma,\mathbf{z}}(f)$ is reached. Now, observe that F is continuous because of continuity of regularization algorithm and

$$\lim_{\gamma\to0}F(\gamma)=f_{\rho,0,\mathbf{z}},\quad\lim_{\gamma\to\infty}F(\gamma)=f_0,\quad f_0\equiv0.$$

Let

$$S_\alpha=\{f\in\mathcal{H}_K:\mathcal{E}_{\rho,0,\mathbf{z}}(f)=\alpha\},\alpha\in[0,\mathcal{E}_{\rho,0,\mathbf{z}}(f_0)).$$

From Proposition 3.2 follows, that $f_{\rho,0,\mathbf{z}} \in S_0$, and from continuity of F, that for each other S_α belongs value of F. Thus, for every α, S_α is non-empty.

Assume that $F(\gamma_\alpha) \in S_\alpha$. Than, $F(\gamma_\alpha)$ minimizes $\|\cdot\|_K$ on S_α, because if not $F(\gamma_\alpha)$ wouldn't be the only infimum $\mathcal{E}_{\rho,\gamma_\alpha,\mathbf{z}}$. It also follows, that there is exactly one value of F in S_α.

Let $g \in S_0$, and let $G : (0,\varepsilon) \to \mathcal{H}_K$, be continuous and let $G(\alpha) \in S_\alpha$ and $\lim_{\alpha\to 0} G(\alpha) = g$. Then, for each $\alpha \in (0,\varepsilon)$, we have $\|F(\gamma_\alpha)\|_K \leq \|G(\alpha)\|_K$, so

$$\|f_{\rho,0,\mathbf{z}}\|_K = \lim_{\alpha\to 0} \|F(\gamma_\alpha)\|_K \leq \lim_{\alpha\to 0} \|G(\alpha)\|_K = \|g\|_K.$$

Hence, we obtain $\|f_{\rho,0,\mathbf{z}}\|_K = \inf_{f\in S_0} \|f\|_K$.

The only thing left, is to show that $f_{\rho,0,\mathbf{z}}$ is the only function for which infimum is reached. Let $f \in S_0 - f_{\rho,0,\mathbf{z}}$. Then, $f + f_{\rho,0,\mathbf{z}} \in S_0$, and

$$0 = \mathcal{E}_{\rho,0,\mathbf{z}}(f + f_{\rho,0,\mathbf{z}}) = \sum_{i=1}^{m} (f(x_i) + f_{\rho,0,\mathbf{z}}(x_i) - y_i)^2 \rho_i = \sum_{i=1}^{m} f(x_i)^2 \rho_i.$$

Consequently, $f(x_i) = 0$, and $S_0 - f_{\rho,0,\mathbf{z}}$ is a subspace of $\mathcal{H}_K$. $S_0 - f_{\rho,0,\mathbf{z}}$ is closed because it has a finite co-dimension. S_0 is a closed affine subspace of $\mathcal{H}_K$, so there is one and only one point in S_0 that minimizes the norm.

4 Approximate Regularization Algorithm

The RA defined above requires solving of system of m linear equations. This is computationally expensive for large m. Instead of solving such linear system exactly, approximate solution can be found. The idea of an approximate algorithm is based on expanding the operator $(\gamma R^{-1} + K[\mathbf{x}])^{-1}$ into series and taking some first terms of this series.

We will use the standard formula for geometric series:

Lemma 4.1. *For $\lambda \in (0,2)$ the following formulas are true:*

$$\sum_{i=0}^{\infty} (1-\lambda)^i = \frac{1}{\lambda}, \quad \sum_{i=0}^{N} (I-\lambda)^i = (I - (I-\lambda)^{N+1})\lambda^{-1}.$$

The formulas are also true for positive operators:

Lemma 4.2. *Let A be positive on a finite dimensional linear space such that all its eigenvalues are in $(0,2)$. Then*

$$\sum_{i=0}^{\infty} (I-A)^i = A^{-1}, \quad \sum_{i=0}^{N} (I-A)^i = (I - (I-A)^{N+1})A^{-1}.$$

Now, we use the formula for operator $\gamma R^{-1} + K[\mathbf{x}]$. Let $\lambda_{\max}$ and $\lambda_{\min}$ be the greatest and smallest eigenvalues of $K[\mathbf{x}]$. Let $\rho_{\max}$ i $\rho_{\min}$ be the greatest and smallest eigenvalues of R.

Theorem 4.3. *Let* $\mu > \frac{1}{2}(\lambda_{\max} + \gamma\frac{1}{\rho_{\min}})$. *Then*

$$(\gamma R^{-1} + K[\mathbf{x}])^{-1} = \frac{1}{\mu}\sum_{i=0}^{\infty}(I - \frac{1}{\mu}(\gamma R^{-1} + K[\mathbf{x}]))^i.$$

Let $K_{\mu,N}$ be a partial sum of series:

$$K_{\mu,N} = \frac{1}{\mu}\sum_{i=0}^{N}(I - \frac{1}{\mu}(\gamma R^{-1} + K[\mathbf{x}]))^i =$$

$$= (I - (I - \frac{1}{\mu}(\gamma R^{-1} + K[\mathbf{x}]))^{N+1})(\gamma R^{-1} + K[\mathbf{x}])^{-1}.$$

We will find the error of approximation after getting partial sum of series. We will assume about μ only that $\mu > 0$. The condition $\mu > \frac{1}{2}(\lambda_{\max} + \gamma\frac{1}{\rho_{\min}})$ is not needed for $K_{\mu,N}$ to be well defined.

Theorem 4.4. *Let* $h(\lambda) = \frac{(1-\frac{\lambda}{\mu})^{N+1}}{\lambda}$, *then*

$$\|(\gamma R^{-1} + K[\mathbf{x}])^{-1} - K_{\mu,N}\|_2 \leq \max(|h(\lambda_{\min} + \gamma\frac{1}{\rho_{\max}})|, |h(\lambda_{\max} + \gamma\frac{1}{\rho_{\min}})|).$$

Proof.

$$(\gamma R^{-1} + K[\mathbf{x}])^{-1} - K_{\mu,N} = (I - \frac{1}{\mu}(\gamma R^{-1} + K[\mathbf{x}]))^{N+1}(\gamma R^{-1} + K[\mathbf{x}])^{-1}.$$

Let Λ be the set of eigenvalues of $\gamma R^{-1} + K[\mathbf{x}]$. Then

$$\|(I - \frac{1}{\mu}(\gamma R^{-1} + K[\mathbf{x}]))^{N+1}(\gamma R^{-1} + K[\mathbf{x}])^{-1}\| = \max_{\lambda\in\Lambda}|(1 - \frac{1}{\mu}\lambda)^{N+1}\lambda^{-1}|.$$

We find the extrema of function h. We have

$$h'(\lambda) = \frac{-(N+1)\frac{\lambda}{\mu}(1-\frac{\lambda}{\mu})^N - (1-\frac{\lambda}{\mu})^{N+1}}{\lambda^2},$$

$h'(\lambda) = 0$ iff $1 - \frac{\lambda}{\mu} = 0$ or $-(N+1)\frac{\lambda}{\mu} - (1 - \frac{\lambda}{\mu}) = 0$. The second equation is identical to $N\frac{\lambda}{\mu} = -1$. Hence, $\lambda < 0$ so it is unimportant to us. The only case left is $1 - \frac{\lambda}{\mu} = 0$, i.e., $\mu = \lambda$.

The function $|h(\lambda)|$ is equal to 0 for $\mu = \lambda$ and is monotonic on intervals $(0, \mu)$ and (μ, ∞). Thus

$$\max_{\lambda\in\Lambda}|h(\lambda)| = \max(|h(\min\lambda \in \Lambda\lambda)|, |h(\max\lambda \in \Lambda\lambda)|) \leq$$

$$\leq \max(|h(\lambda_{\min} + \gamma\frac{1}{\rho_{\max}})|, |h(\lambda_{\max} + \gamma\frac{1}{\rho_{\min}})|).$$

Now, we can find the μ that minimizes the error. Let

$$c = \frac{\lambda_{\max} + \gamma \frac{1}{\rho_{\min}}}{\lambda_{\min} + \gamma \frac{1}{\rho_{\max}}}.$$

The coefficient c is called the *condition* of operator $\gamma R^{-1} + K[\mathbf{x}]$.

Theorem 4.5. *The error is minimized by*

$$\mu = \big(\lambda_{\min} + \gamma \frac{1}{\rho_{\max}}\big) \frac{\sqrt[N+1]{c}}{\sqrt[N+1]{c} + 1},$$

and is equal to

$$\frac{1}{\lambda_{\min} + \gamma \frac{1}{\rho_{\max}}} \big(\frac{c-1}{c + \sqrt[N+1]{c}}\big)^{N+1}.$$

Proof. Let $\lambda_1 = \lambda_{\min} + \gamma \frac{1}{\rho_{\max}}$ and $\lambda_2 = \lambda_{\max} + \gamma \frac{1}{\rho_{\min}}$.
From the shape of function $|h|$ and the error formula in the above theorem it follows that we seek for μ satisfying the following conditions:

$$\lambda_1 < \mu < \lambda_2 \quad \text{and} \quad |h(\lambda_1)| = |h(\lambda_2)|.$$

From the definition of h we obtain

$$|h(\lambda)| = \Big|\frac{(1 - \frac{\lambda}{\mu})^{N+1}}{\lambda}\Big| = \big(\frac{|1 - \frac{\lambda}{\mu}|}{\sqrt[N+1]{\lambda}}\big)^{N+1}.$$

So, we look for μ such that:

$$\frac{|1 - \frac{\lambda_1}{\mu}|}{\sqrt[N+1]{\lambda_1}} = \frac{|1 - \frac{\lambda_2}{\mu}|}{\sqrt[N+1]{\lambda_2}},$$

$$(\mu - \lambda_1) \sqrt[N+1]{c} = \lambda_2 - \mu,$$

$$\mu(\sqrt[N+1]{c} + 1) = \lambda_2 + \lambda_1 \sqrt[N+1]{c} = \lambda_1(c + \sqrt[N+1]{c}).$$

From this we obtain μ and the error can be simply calculated:

$$\frac{(1 - \frac{\lambda_1}{\mu})^{N+1}}{\lambda_1} = \frac{1}{\lambda_1}\big(1 - \frac{\sqrt[N+1]{c} + 1}{c + \sqrt[N+1]{c}}\big)^{N+1} = \frac{1}{\lambda_1}\big(\frac{c-1}{c + \sqrt[N+1]{c}}\big)^{N+1}.$$

Now, we define the iterative algorithm that calculates $K_{\mu,N}\mathbf{y}$ by means of N multiplications of $K[\mathbf{x}]$ by a vector. Let

$$\mathbf{y}_0 = \mathbf{y},$$

$$\mathbf{y}_{N+1} = \frac{1}{\mu}\mathbf{y} + (I - \frac{1}{\mu}(\gamma R^{-1} + K[\mathbf{x}]))\mathbf{y}_N.$$

Theorem 4.6.

$$\mathbf{y}_N = K_{\mu,N}\mathbf{y}.$$

Proof. Let

$$K_{\mu,0} = I,$$

$$K_{\mu,N+1} = \frac{1}{\mu}\sum_{i=0}^{N+1}(I - \frac{1}{\mu}(\gamma R^{-1} + K[\mathbf{x}]))^i = \frac{1}{\mu}I + (I - \frac{1}{\mu}(\gamma R^{-1} + K[\mathbf{x}]))K_{\mu,N}.$$

So

$$\mathbf{y}_0 = K_{\mu,0}\mathbf{y},$$

$$\mathbf{y}_{N+1} = \frac{1}{\mu}I\mathbf{y} + (I - \frac{1}{\mu}(\gamma R^{-1} + K[\mathbf{x}]))K_{\mu,N}\mathbf{y} = K_{\mu,N+1}\mathbf{y}.$$

5 Conclusions

There are many algorithms for solving of the learning-approximation problem [5], [9]. The algorithms differ in properties, complexity, hypothesis spaces. Yet it is impossible to judge which one is the best. The properties distinguishing the Regularization Algorithm make it possible to create approximation that preserve the values for known samples and is smooth at the same time. The other advantage is the firm mathematical theory that allows a deep analysis of the algorithm behavior.

That's why the good applications for regularization algorithm seems to be the ones in which samples have low error rate and the details are important.

The meaning of the condition of the $K[\mathbf{x}]$ matrix remains an open question. The condition plays crucial role for the error value of the approximate RA. It also seems to be related to the dimension of the example space and might be an important statistics for determining such a dimension.

We plan to use the RA algorithm for inducing local models for each classified new object. For any new object first the set of objects relevant for such object is extracted from data (e.g., using knn strategy [5]) and next the RA algorithm can be applied to the extracted data. Such an approach seems to follow the main idea outlined in Epilog of [9]

In our future study we also would like to develop a "discrete" version of the RA algorithm assuming that objects can be indiscernible, i.e., assuming that reasoning is performed under granularity of knowledge [6], [7].

Acknowledgements

The research has been supported by the grant 3T11C00226 from Ministry of Scientific Research and Information Technology of the Republic of Poland.

References

1. F. Cucker and S. Smale , *On the mathematical foundations of learning*, Bulletin of AMS, 39:1-49, 2001.

2. F. Cucker and S. Smale, *Best choices for regularization parameters in learning theory*, Foundations of computational Mathematics, 2(4):413-428, 2002.
3. T. Dodd, R. Harrison, *Iterative Solution to Approximation in Reproducing Kernel Hilbert Spaces*, 15th IFAC World Congress: b'02 (CDROM).
4. T. Evgeniou, M. Pontil and T. Poggio, *Regularization Networks and Support Vector Machines*, Advances in Computational Mathematics 13 (2000) 1-50.
5. J. H. Friedman, T. Hastie, R. Tibshirani, *Statistical Learning: Data Mining, Inference, and Prediction*, Springer-Verlag, Heidelberg, 2001
6. S. K. Pal, L. Polkowski, A. Skowron (Eds.), *Rough-Neural Computing: Techniques for Computing with Words*, Cognitive Technologies. Springer-Verlag, 2004.
7. Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning about Data*, Kluwer Academic Publishers, Dordrecht, 1991.
8. T. Poggio and S. Smale, *The Mathematics of Learning: dealing with Data*, Notices of the AMS, 50(5):537-544, 2003.
9. V. N. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.