6

Recent advances in Neural Networks Structural Risk Minimization based on multiobjective complexity control algorithms

D.A.G. Vieira, J.A. Vasconcelos and R.R. Saldanha
Department of Electrical Engineering
Federal University of Minas Gerais
Brazil

Nowadays, neural networks (NNs) are widely applied in the solution of several real world problems. They have been successfully used in many fields such as chemistry, physics, engineering, and bio-informatics among others. However, their use often relies on some handcrafted settings, such as the number of layers and neurons. This chapter will discuss the Structural Risk Minimization (SRM) problem using some multiobjective optimization concepts. Both are closely related to the classical Tikhonov's regularization scheme, and, it is also exploited in this work.

A neural network is a learning machine capable to describe, to the input x, the set of functions $\mathbb{F} = \{f(x,w) : x \in \mathbb{X},\ w \in \mathbb{W}\}$, where $\mathbb{W}$ is the space of possible weights. Given a supervisor which defines an output vector $y \in \mathbb{Y}$ (desired output), for a given input x, according to the conditional distribution $F(y|x)$, the ultimate goal in the learning problem is to find $w \in \mathbb{W}$ that best approximates the supervisor answer given some measure. To some loss function $L(.)$, the expected risk (error) can be defined as, Vapnik (1998):

$$R(w) = \int L(y, f(x,w))\, dF(x,y). \tag{1}$$

Therefore, the learning problem can be understood as finding $f(x,w_0) : w_0 \in \mathbb{W}$, such that $R(w)$ is minimal. Nevertheless, the function $F(x,y)$ is unknown, thus, it is impossible to direct evaluate $R(w)$. The only available information about the supervisor is contained in the training set $\mathbb{T} = \{(x_1,\tilde{y}_1),...,(x_t,\tilde{y}_t)\}$. Where $\tilde{y}$ is y plus some uncertainty, as noise. For instance, for regression and prediction problems $y \in \mathbb{R}^t$, and for binary classification $y \in \{-1,1\}^t$.

In the early years of NN research, it was believed that decreasing the training error (empirical risk) was a sufficient condition to approximate the supervisor answer. This problem was stated as

$$w_* = \arg\min_{w \in \mathbb{W}} J(\mathbb{T},w) = \frac{1}{t}\sum_{i=1}^{t} L(\tilde{y}_i, f(x_i,w)). \tag{2}$$

This approach was considered self-evident for many years and the main milestone was to find better algorithms to solve (2). However, the non self-evident overfitting phenomenon has appeared. This would imply that $w_* \neq w_0$. One way to characterize it is by the bias and

variance dilemma, S. Geman & Doursat (1992). The expected mean-squared error between $f(\cdot)$ and the expected value of y given x, $E[y|x]$, can be written as:

$$E_{\mathbb{T}}[(f(x;\mathbb{T}) - E[y|x])^2] = \begin{array}{l}(E_{\mathbb{T}}[f(x;\mathbb{T})] - E[y|x])^2 \\ +E_{\mathbb{T}}[(f(x;\mathbb{T}) - E_{\mathbb{T}}[f(x;\mathbb{T})])^2],\end{array} \tag{3}$$

where $E_{\mathbb{T}}[.]$ is the expected value given a set $\mathbb{T}$. The first term in the right hand side of (3) is known as bias, and the second one as variance. The variance term measures the sensibility of the approximating function given a data set $\mathbb{T}$. To control the variance, models with less complexity should be generated, i.e., they cannot change too much to a given data $\mathbb{T}$. On the other hand, some bias is inserted in the problem when the complexity is limited, thus, this should be controlled.

This chapter is organized as follows. First, the regularization theory from Tikhonov (1963), a well-known technique to solve linear ill-posed problems, will be introduced together with the residual method from Phillips (1962) and the quasi-solutions from Ivanov (1962; 1976). It is shown, using Singular Value Decomposition (SVD), the relationship between these methods and the Wiener's filter. After that, the Structural Risk Minimization (SRM), and the multiobjective learning will be discussed. These methods are closely related, and, some of their main aspects will be discussed. Inspired on the Tikhonov's regularization it will be discussed the well-known weight decay (WD) method for NNs, Hinton (1989). However, it will be clarified that this method is not consistent if the functions are not convex, which is usually the case. To overcome that, it is introduced the generalized Tikhonov's regularization based on a Q-norm for Parallel Layers Perceptrons (PLPs). Finally, some results are presented.

1. Linear ill-posed problems

Given the linear mapping $A : \mathbb{W} \mapsto \mathbb{Y}$, the equation

$$Aw = y,\ A \in \mathbb{R}^{t\times n},\ w \in \mathbb{R}^n \ and\ y \in \mathbb{R}^t, \tag{4}$$

is well-posed provided that: (i) for each $y \in \mathbb{Y}$, $\exists w \in \mathbb{W}$ such that $Aw = y$ (existence); (ii) $Aw_1 = Aw_2 \Leftrightarrow w_1 = w_2$ (uniqueness); (iii) A^{-1} is continuous (stability). Thus, a problem is called well-posed if its solution exists, is unique and stable. Unfortunately, inverse problems, such as the ones to select a model based on the data, are usually ill-posed, i.e., it violates at least one of the aforementioned conditions. In applied sciences and engineering the right-hand side vector y can be contaminated by noise, $\xi \in \mathbb{R}$, thus, instead of y only $\tilde{y}$ is available and

$$\|y - \tilde{y}\|_2 \leqslant \xi. \tag{5}$$

The problem is said to be stable if small variations in the right-hand side implies small changes in the solution

$$\|w - \tilde{w}\|_2 \leqslant \delta(\xi). \tag{6}$$

The existence can be imposed by considering the minimal Euclidian norm

$$w_* = \arg\min_{w\in\mathbb{W}} J(w) = \|Aw - \tilde{y}\|_2^2 = (Aw - \tilde{y})^T(Aw - \tilde{y}). \tag{7}$$

Making $\partial J/\partial w = 0$[1],

$$w_* := (A^T A)^{-1} A^T \tilde{y} := A^{\dagger} \tilde{y}, \tag{8}$$

where $A^{\dagger}$ is the pseudo inverse. Considering w_0 the desired solution of (4) and w_* the solution of (7), due to the error in y and the ill-conditioning in A, the following relation is usually true

$$\|w_*\| \gg \|w_0\|, \tag{9}$$

which is not a meaningful approximation of w_0. In the early 60's, Tikhonov proposed the regularization method to solve this problem, Tikhonov (1963). The Tikhonov's method considers the solution of the following auxiliary problem:

$$w_\lambda = \arg\min_{w \in \mathbb{W}} J_\lambda = \|Aw - \tilde{y}\|_2^2 + \lambda \Omega(w), \tag{10}$$

where $\lambda > 0$ is a pre-defined constant known as regularization parameter. The regularization function $\Omega(w)$ is defined as semi-continuous, positive and compact in the space of functions defined by w, i.e., $\Omega(w) \leq c$, $c > 0$. To guarantee the uniqueness of the solution the following properties are required: (i) $\Omega(w)$ is a non-negative convex function; (ii) $\Omega(0) = 0$ holds true; and (iii) the $r(\rho) = \Omega(\rho w)$ is strictly growing function. This method is usually written as

$$\begin{aligned} w_\lambda = \arg\min_{w \in \mathbb{W}} J_\lambda \quad &= \|Aw - \tilde{y}\|_2^2 + \lambda \|w\|_2^2 \\ &= (Aw - \tilde{y})^T (Aw - \tilde{y}) + \lambda x^T I x \end{aligned} . \tag{11}$$

For each positive parameter λ, considering the complexity $\Omega = w^T I w = \|w\|_2^2$, where I is the identity matrix, (10) has a unique solution of the following form:

$$w_\lambda := (A^T A + \lambda I)^{-1} A^T \tilde{y}. \tag{12}$$

This result was fundamental to the popularization of the Tikhonov's technique, since it has a simple closed-form solution. In statistics it is also known as ridge regression. In fact, Tikhonov & Arsenin (1977) proved that w_λ converges to w_0 as $\xi \to 0$ if

$$\lim_{\xi \to 0} \lambda(\xi) = 0, \tag{13}$$

$$\lim_{\xi \to 0} \frac{\xi^2}{\lambda(\xi)} = 0. \tag{14}$$

Consider the set $\mathbb{W}_k = \{w : \Omega(w) \leq c_k\}$, $c_k > 0$. Since Ω defines a compact subset the following holds true

$$\mathbb{W}_1 \subseteq \mathbb{W}_2 \subseteq ... \subseteq \mathbb{W}_i, ... \Rightarrow c_1 < c_2 < ... < c_i, ... \tag{15}$$

Define w_{k*} as

$$w_{k*} = \arg\min_{w \in \mathbb{W}_k} \|Aw - \tilde{y}\|. \tag{16}$$

[1] Using:

$$\frac{\partial(\tilde{y}^T w)}{\partial w} = \tilde{y}, \ \frac{\partial(w^T A^T A w)}{\partial w} = (A^T A + A^T A) w$$

For some general conditions, Ivanov (1962; 1976) proved that the sequence $w_{1*},...,w_{k*}$ converges to w_0, the desired solution. This is called quasi-solutions method and can be written, for some $\epsilon > 0$, as

$$\begin{aligned} w_\epsilon = \arg\min_{w \in \mathbb{W}} \|Aw - \tilde{y}\|_2^2 \\ \text{subject to: } \|w\|_2^2 \leq \epsilon \end{aligned} \tag{17}$$

In the same period Phillips (1962) proposed the residual method

$$\begin{aligned} w_\epsilon = \arg\min_{w \in \mathbb{W}} \|w\|_2^2 \\ \text{subject to: } \|Aw - \tilde{y}\|_2^2 \leq \epsilon \end{aligned}. \tag{18}$$

In Vasin (1970) it is shown that the Regularization, Residual and Quasi-solutions methods are equivalent, i.e., they can generate the same set of solutions, given the linear problem stated in (4), and the distance measured using the Euclidian norm. Consider the problem stated in Alavetti & Eichel (2004)

$$\begin{aligned} w_\Delta = \arg\min_{w \in \mathbb{W}} \|Aw - \tilde{y}\|_2^2 \\ \text{subject to: } \|w\|_2^2 = \Delta \end{aligned}. \tag{19}$$

Assuming that $\|w_0\| > \Delta$, this constrained minimization problem has a unique solution $w_{\lambda,\Delta}$ of the form (12). The value of λ is positive such that $\|w_\lambda\| = \Delta$ Alavetti & Eichel (2004). Assume that $A^\dagger \tilde{y} \neq 0$, the function

$$\varphi(\lambda) := \|w\|^2, \ \ \lambda \geq 0, \tag{20}$$

can be expressed as

$$\varphi(\lambda) := \tilde{y} A (A^T A + \lambda I)^{-2} A^T \tilde{y}, \ \ \lambda > 0, \tag{21}$$

which shows that $\varphi(\lambda)$ is strictly decreasing and convex for any $\lambda > 0$, and that, $\varphi(\lambda) = \Delta$ has a unique solution λ, such that $0 < \lambda < \infty$, for any Δ that satisfies $0 < \Delta < \|A^\dagger \tilde{y}\|^2$, Alavetti & Eichel (2004).
Even though all the results considered so far used the Euclidian norm $\|.\|_2$ to define the complexity Ω, the more general p-norm

$$\|w\|_p = \sum_{i=1}^{n} |w_i|^p, \tag{22}$$

can also be applied. This is the case of the shrinkage method called Lasso, Hastie et al. (2001)

$$\begin{aligned} w_{lasso} = \arg\min_{w \in \mathbb{W}} \|Aw - \tilde{y}\|_2^2 \\ \text{subject to: } \|w\|_1 \leq \epsilon \end{aligned}. \tag{23}$$

This chapter will concentrate in the Euclidian norm based formulation due to their simplicity, and the existence of closed form solutions. According to Hastie et al. (2001) it could be used any p besides 1, or 2, and that, indeed, we could try to estimate it from the data, but there is no results in this direction so far.

1.1 Wiener's filter interpretation

Consider the singular value decomposition (SVD) of A as

$$A = USV^T \tag{24}$$

where U and V are unitary matrices, i.e, $U^{-1} = U^T$, and $S = diag(s_1, s_2, ..., s_t)$ is a diagonal matrix with $s_1 \geq s_2 \geq ... \geq s_t \geq 0$, called the singular values of A. Thus, w_λ, given in (12), can be written as:

$$\begin{aligned} w_\lambda &= (VS^TU^TUSV^T + \lambda VIV^T)^{-1}VS^TU^T\tilde{y} \\ &= V(S^TS + \lambda I)^{-1}S^TU^T\tilde{y} \end{aligned}$$

$$w_\lambda = \sum_{i=1}^{t} \frac{s_i^2}{s_i^2 + \lambda} \frac{u_i^T\tilde{y}}{s_i} v_i, \tag{25}$$

where $\lambda \geq 0 : \frac{s_i^2}{s_i^2+\lambda} \leq 1$ are the Wiener's filter weights. The SVD of the matrix A is related to the principal component analysis. Therefore, it implies that it shrinks more the directions with smaller variance. Next section will introduce the Weight Decay, the realization of the ideas presented in this section to Neural Networks.

2. Structural Risk Minimization principle

The structural risk minimization (SRM) was introduced by Vapnik and Chervonenkis and a description of it can be found Vapnik (1992), Vapnik (1998). One of the main achivements of the SRM is the introduction of the idea of capacity of a set of functions. It is based on some theoretical results that shows that the upper bound of the learning machine expected risk depends on: (i) the training error and, (ii) the machine capacity, defined as the VC dimension and its variations, Vapnik (2001). This inductive principle is directly applied in learning machines as the Support Vector Machines (SVMs). Following these considerations the SRM principle considers the minimization of two factors: the training error and the VC dimension.
Consider the function $J(\cdot,\cdot) : \mathbb{Z} \times \mathbb{W} \mapsto \mathbb{R}$, in which $\mathbb{Z}$ and $\mathbb{W}$ are arbitrary spaces. Taking its second argument $w \in \mathbb{W}$ as a parameter constrained to a set $\mathbb{W}_k \subset \mathbb{W}$, a set $\mathbb{J}$ of functions $J(\cdot, w) : \mathbb{Z} \mapsto \mathbb{R}$ becomes defined for $w \in \mathbb{W}_k$. This set can be structured as a sequence of nested subsets $\mathbb{J}_k = \{J(\cdot, w),\ w \in \mathbb{W}_k\}$, such that

$$\mathbb{W}_1 \subset \mathbb{W}_2 \subset ... \subset \mathbb{W}_i ... \Rightarrow \mathbb{J}_1 \subseteq \mathbb{J}_2 \subseteq ... \subseteq \mathbb{J}_i ... \tag{26}$$

The sequence (26) should fulfill the following conditions: (i) the VC dimension, h_k, of each set $\mathbb{J}_k$ is finite, and (ii)

$$h_1 \leq h_2 \leq ... \leq h_i ... \,. \tag{27}$$

For any positive integer k, there is a finite positive scalar B_k such that $J(z, w) \leq B_k$, $\forall\ w \in \mathbb{W}_k$ and $z \in \mathbb{Z}$. The principle of SRM is oriented to find the values of w and k such that $w \in \mathbb{W}_k$, making the function $J(\cdot, w)$ minimize the empirical risk, while the set $\mathbb{W}_k$ minimizes the structural risk.

2.1 Multiobjective Learning

The SRM can be interpreted as a bi-objective optimization problem, which considers the minimization of the empirical risk and the machine capacity. Instead of the integer index k, a straightforward generalization is to consider that the set $\mathbb{W}$ is parameterized by a continuous parameter ζ . Given a training set $\mathbb{T}$, the SRM problem for this set can be written as:

$$\text{(SRM):} \quad \min_{\zeta,w} \begin{cases} J(\zeta,w) \\ \Omega(\zeta,w) \end{cases} . \tag{28}$$

in which J represents some empirical risk function, and Ω the complexity of the learning machine, for instance the fat-shattering dimension, Shawe-Taylor & Bartlett (1998).
Usually, it is not possible to minimize J and Ω simultaneously, because the optimum to one function hardly ever is the optimum to the other one. Thus, there is not a single optimum, but a set of them, when a multiobjective formulation is considered. In order to state the solutions of the SRM, the following definitions are required:

(i) *Dominance*: A pair (ζ_a, w_a) dominates another pair (ζ_b, w_b), which is denoted by $(\zeta_a, w_a) \prec (\zeta_b, w_b)$, if $J((\zeta_a, w_a)) \leq J((\zeta_b, w_b))$ and $\Omega((\zeta_a, w_a)) \leq \Omega((\zeta_b, w_b))$, with the strict inequality valid for at least one of the functions.

(ii) *Pareto optimality*: A pair (ζ_*, w_*) is called Pareto-Optimal (PO) if there is no other feasible pair which dominates it.

By using these definitions, it is possible to generate the set of solutions called PO front, which have the best trade-off between the error and the machine complexity. All such solutions are candidate solutions for the SRM problem.
Examining (26) and (27) from the viewpoint of the Pareto Optimality of (44), it can be seen that in the nested sequence $\mathbb{J}_1 \subset \mathbb{J}_2 \subset ... \subset \mathbb{J}_i ...$, the minimal empirical error in the set is ordered as $J_{1*} \geq J_{2*} \geq ... \geq J_{i*}$, where $J_{k*} := J(w_{k*})$ and

$$\begin{aligned} & w_{k*} = \arg\min_{w} J(w) \\ & \text{subject to: } \; w \in \mathbb{W}_k \end{aligned} \tag{29}$$

The solutions w_{k*} are Pareto-Optimal ones, each one associated to the corresponding sequence set $\mathbb{J}_k$. These are the solutions of the SRM problem. Any other function $J(\cdot, w)$ that is not a solution of any minimization problem of this form must be dominated, and cannot be a solution of the SRM problem. This will be the base of some novel results presented in this chapter. Defining the complexity as $\Omega(\zeta)$, it can be associated to some $\mathbb{W}(\zeta)$ defined by

$$\mathbb{W}(\zeta) = \{w : \; \|w\| < \zeta\} \tag{30}$$

and

$$\Omega(\zeta) = \zeta. \tag{31}$$

Given $\zeta_1 < \zeta_2$, this choice of $\mathbb{W}(\zeta)$ and $\Omega(\zeta)$ preserves the necessary relations:

- $\mathbb{W}(\zeta_1) \subset \mathbb{W}(\zeta_2)$;
- $J(\cdot, \zeta_1) \geq J(\cdot, \zeta_2)$;
- $\Omega(\zeta_1) < \Omega(\zeta_2)$.

As the minimization of the structural risk $\Omega(\zeta) = \zeta$ is equivalent to the minimization of the norm of w, the structural risk minimization principle becomes, in this case, stated in terms of w only:

$$\text{(SRM):} \quad \min_{w} \left\{ \begin{array}{l} J(\cdot, w) \\ \Omega(w) = \|w\| \end{array} \right. . \tag{32}$$

3. The Weight Decay for MLPs

The Multi-Layer Perceptron (MLPs) is a popular neural network which considers the neurons (or perceptrons) in cascade. Consider the input vector x, which includes the bias term, i.e., it is added an extra element equal to 1, the vectorial function Φ, and the weight matrix W

$$\begin{array}{ll} \Phi_1 & = \phi_1(W_1^T x) \\ & \downarrow \\ \Phi_q & = \phi_q(W_q^T \Phi_{q-1}) \end{array} , \tag{33}$$

then

$$f(x,w)^{MLP} = \Phi_q(\Phi_{q-1}(...\Phi_1(.))) \tag{34}$$

where q is the number of layers, and ϕ is an activation function as hyperbolic tangent. For a weight matrix W and the vector w_j

$$w_j^T x = \sum_{i=0}^{n} w_{ji} x_i. \tag{35}$$

Therefore, the MLPs implement a nonlinear function of the sum of nonlinear functions. With one hidden layer, and m neurons, it can be written as:

$$f(x,w)^{MLP} = \sum_{i=1}^{m} W_{2i} \phi(W_{1i}^T x), \tag{36}$$

where $x \in \mathbb{R}^{n+1}$, $x_0 = 1$ is the bias, W_2 is a vector with m elements and W_1 is a matrix $((n+1) \times m)$. The vector w is defined as a vector which contains all the elements of W_i.
Using the ideas from the regularization framework, the Weight Decay (WD) is a direct implementation of the Tikhonov's model to MLPs. The WD consists in writing a weighted sum of the Empirical risk, $J(\cdot)$, and the norm of the weight vector

$$w_\lambda = \arg\min_{w} J_\lambda^{MLP} = J(w,x) + \lambda \|w\|_2^2, \tag{37}$$

where $J(\cdot)$

$$J(w,x) = \frac{1}{2} \sum_{i=1}^{t} (f(x_i,w) - \tilde{y}_i)^2. \tag{38}$$

In Bartlett (1998) it was shown that the fat-shattering dimension, which is a generalization of the VC dimension, can be limited by limiting the weights of a given network. Limiting the fat-shattering dimension leads to a limit in the generalization error, Vapnik (1998; 2001). This gives support to the use of the norm of the weight vector as the complexity constraint. The main difference between the problem stated in (10) and (37) is that in the first one the risk is guaranteed to be convex, while in the second one it can be non-convex and even multi-modal. Next section will show that the weighted sum approach, which is the base of the WD method, is not consistent given non-convex problems.

3.1 The convexity issue

The WD approach is based on the general weighted sum function

$$J_\lambda(w) = \lambda J_1 + (1-\lambda)J_2, \tag{39}$$

where $\lambda = [0,1]$ controls the importance of the objectives. Consider the following non-convex unimodal one-variable functions:

$$J_1(w) = ((w-1)^2 - \tanh(40w-4))^2, \tag{40}$$

$$J_2(w) = 200w^2, \tag{41}$$

where the factor 200 was used only to simplify presentation. Given, $\lambda = 0.3$ and $\lambda = 0.6$, the following weighted sum functions can be written

$$J_a(w) = 0.3J_1 + (1-0.3)J_2, \tag{42}$$

$$J_b(w) = 0.6J_1 + (1-0.6)J_2. \tag{43}$$

The functions J_1 and J_2 and two possible weighted solutions, J_a and J_b are shown in Fig. 1. Note that the weighted functions have become multimodal, although the original functions were unimodal. The PO front for this problem is presented in Fig. 2 and it is composed of both convex and non-convex parts.

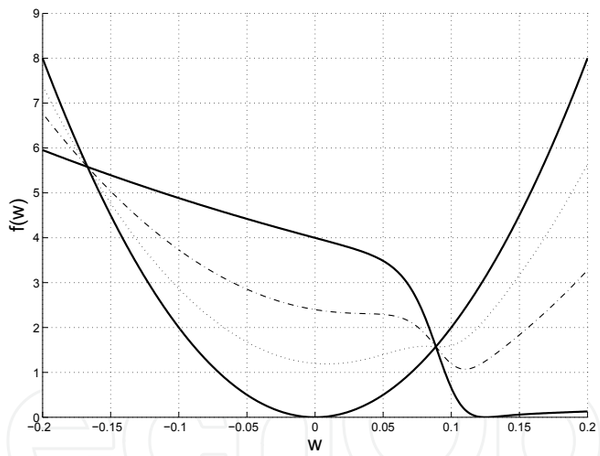

Fig. 1. The original functions are presented in continuous line$(-)$. Two possible weighted solutions for this problem, with $\lambda = 0.3$ and $\lambda = 0.6$ are shown in $(-.)$ and $(--)$, respectively.

The relevant conclusion here is: if J_1 and J_2 are not convex functions (what is the case in most of machine learning problems), the weighed sum approach should not be employed for trying to find the trade-off front, as it may loose some potential solution.

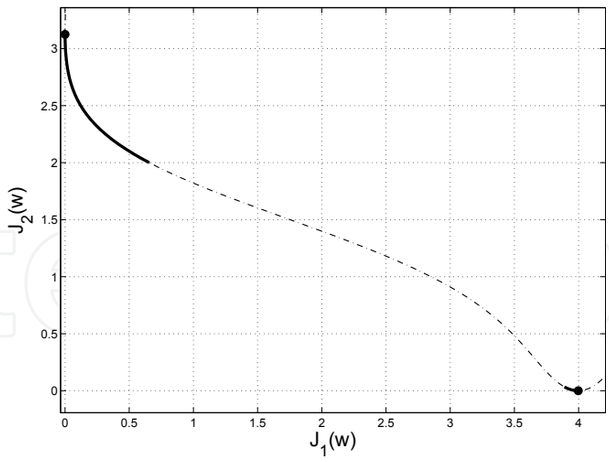

Fig. 2. The minima of J_1 and J_2 are marked with the •, being the PO front everything between them. The convex part of the PO front is marked using continuous lines (-) and the non-convex as (–). The WD method can only generate the networks which belong to the convex part.

4. The Parallel Layer Percetron

Instead of assembling the layers in cascade, in Caminhas et al. (2003) it was proposed to use them in parallel, given birth to the Parallel Layer Percetron (PLP). Consider the input vector x, which includes the bias term, the vector function Φ, and the weight matrix W

$$\begin{array}{ll} \Phi_1 & = \phi_1(W_1^T x) \\ & \downarrow \\ \Phi_q & = \phi_q(W_q^T x) \end{array}, \tag{44}$$

then

$$f(x,w)^{PLP} = \phi\left(\prod_{i=1}^{q}\Phi_i\right), \tag{45}$$

where, $\prod$ represents a point wise product, and w is a vector with all the weights W_i. Hence, the PLP implements a nonlinear function of the product of nonlinear functions. This configuration has some computational advantages as discussed in Caminhas et al. (2003). A particular case of this topology can be written as the sum of the product of a linear layer, $L^T x$, and a nonlinear layer, $\Phi = \phi(N^T x)$, and it is given by

$$f(x,w)^{PLP} = x^T L\Phi^T = \sum_{j=1}^{m}\left[\sum_{i=0}^{n} L_{ji}x_i\phi(\sum_{i=0}^{n} N_{ji}x_i)\right]. \tag{46}$$

Since $f(x,w)^{PLP}$ is a linear function of the parameters L_{ji}, the PLP output can be written in a matrix form. Thus, consider the vector $l_z = L_{ji}$, where $z = (n+1)(j-1)+i$. This vector is a matrix transformation L to a vector with the same components, where $j = 1,...,m$, $i = 0,\ldots,n$. By calculating all the outputs of the nonlinear perceptrons, a matrix A, with components $a_{kz} =$

$x_{zk}\phi(N_j^T x_i), k = 1,...,t$ can be constructed

$$A = \begin{bmatrix} x_{01}\phi(b_{11}) & \dots & x_{n1}\phi(b_{m1}) \\ \vdots & \dots & \vdots \\ x_{0t}\phi(b_{1t}) & \dots & x_{nt}\phi(b_{mt}) \end{bmatrix}. \tag{47}$$

Therefore, the output of the PLP network can be written as

$$f(x,w)^{PLP} = A(x,N)l. \tag{48}$$

Thus, the empirical risk can be written as:

$$J^{PLP}(\mathbb{T},w) = (Al - \tilde{y})^T(Al - \tilde{y}). \tag{49}$$

In this case the error is a quadratic function of the control variables - the vector l - while A is a nonlinear function of N. The main idea that will follow is to find a formulation, which resembles the Tikhonov's least squares solution, for this topology. Even though it is clear how to use the vector l, the nonlinear weights N brings an additional complication. To solve this problem it is necessary to find a function $\Omega(l)$ which is capable to consider also the complexity derived from N. For that, a generalized version of the Tikhonov's regularization, based on a Q-norm, can be used.

5. Generalized Tikhonov's regularization using a Q-norm

For any norm and any bijective linear transformation D, a new norm of l can be defined to be equal to $\|Dl\|$. For instance, in 2D, with D a rotation by 45 and a suitable scaling, this changes the 1-norm into an ∞-norm. Consider the Euclidean norm of the transformed vector

$$\|Dl\|_2 = \sqrt{l^T D^T Dl} = \sqrt{l^T Ql}, \tag{50}$$

for $Q = D^T D$, $w \in \mathbb{R}^n$ a vector with finite dimension, and $D^T D = Q \in \mathbb{R}^{n\times n}$ a symmetric positive definite matrix, i.e., $l^T Ql > 0$, $\forall l \neq 0$.. The Q-norm of w is given by $\sqrt{l^T Qw}$. The regularization function Ω can be written as a Q-norm, where the matrix Q is a function of the nonlinear parameters N:

$$\Omega(l,N) = l^T Q(N)l. \tag{51}$$

Therefore, the solution of the linear ill-posed problem can be generalized as

$$l_Q = \arg\min_{l\in\mathbb{W}} J_\lambda = \|Al - \tilde{y}\|^2_{Q_1} + \lambda\|l\|^2_{Q_2} \tag{52}$$

Thus, it is need to define a matrix Q_2 such that it considers the influence of the nonlinear parameters of the PLP, while only adjusting the linear ones. This will be achieved using the Minimum Gradient Method (MGM).

5.1 The Minimum Gradient Method

By calculating the derivative of (46) with respect to x_k, one obtains, Vieira et al. (2008):

$$\frac{\partial f(x,w)}{\partial x_k} = \sum_{j=1}^{m} \left[\left(\frac{\partial \phi}{\partial b_j} N_{jk} \right) \left(\sum_{i=0}^{n} L_{ji} x_i \right) + \phi(b_j) L_{jk} \right]. \tag{53}$$

where $b_j = \sum_{i=0}^{n} N_{ji} x_i$. For all j and $z = (n+1)(j-1)+i$ the following holds true

$$\frac{\partial f(x,w)_j}{\partial x_k} = \left[\left(\frac{\partial \phi}{\partial b_j} N_{jk} x_k \right) + \phi(b_j) \right] l_z, \; k = i, \tag{54}$$

$$\frac{\partial f(x,w)_j}{\partial x_k} = \left[\left(\frac{\partial \phi}{\partial b_j} N_{jk} x_i \right) \right] l_z, \; k \neq i. \tag{55}$$

The derivatives in relation to the vector $x_k = [x_{k1},...,x_{kh},...,x_{kt}]^T$, where t is the number of samples, can be written in a vector form as follows:

$$\frac{\partial f(x,w)}{\partial x_k} = D_k l. \tag{56}$$

To exemplify the construction of the matrix D_k, where $D_k \in \mathbb{R}^{t \times (n+1)m}$, consider the following cases when the derivatives in relation to x_1 and x_2 are computed, for b_{jh}, N_{ji}, x_{ih}, where j represents the neuron, i the input and h the sample number.

$$D_1 = \begin{bmatrix} \frac{\partial \phi}{\partial b_{11}} N_{10} & \frac{\partial \phi}{\partial b_{11}} N_{11} x_{11} + \phi(b_{11}) & \cdots & \frac{\partial \phi}{\partial b_{m1}} N_{mn} x_{n1} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial \phi}{\partial b_{1t}} N_{10} & \frac{\partial \phi}{\partial b_{1t}} N_{11} x_{1t} + \phi(b_{1t}) & \cdots & \frac{\partial \phi}{\partial b_{mt}} N_{mn} x_{nt} \end{bmatrix} \tag{57}$$

$$D_2 = \begin{bmatrix} \frac{\partial \phi}{\partial b_{11}} N_{10} & \frac{\partial \phi}{\partial b_{11}} N_{11} x_{11} & \frac{\partial \phi}{\partial b_{11}} N_{12} x_{21} + \phi(b_{11}) & \cdots & \frac{\partial \phi}{\partial b_{m1}} N_{mn} x_{n1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \frac{\partial \phi}{\partial b_{1t}} N_{10} & \frac{\partial \phi}{\partial b_{1t}} N_{11} x_{1t} & \frac{\partial \phi}{\partial b_{1t}} N_{12} x_{2t} + \phi(b_{1t}) & \cdots & \frac{\partial \phi}{\partial b_{mt}} N_{mn} x_{nt} \end{bmatrix} \tag{58}$$

In the matrices D_k, when $i = k$, the columns related to the weights $L_{jk} = l_z$ are composed by two terms, as can be noticed in the second column of D_1 and in the third one of D_2. In the other columns just one term is used. Remembers that $i = 0$ represents the bias term. Therefore, the complexity function Ω can be defined as the minimization of the norm of the output gradient

$$\Omega^{PLP} = \sum_{k=1}^{n} (D_k l)^T (D_k l) = l^T Q l, \tag{59}$$

where $Q \in \mathbb{R}^{(n+1)m \times (n+1)m} = \sum_{k=1}^{n} D_k^T D_k$. Clearly, $l^T Q l \geq 0 \; \forall l$, noticing that the sum of symmetric positive-definite matrices are also symmetric positive-definite matrices. The construction of the matrix $D_k^T D_k$ is exemplified taking $k = 2$:

$$D_2^T D_2 = \begin{bmatrix} \sum_{h=1}^{t} \left(\frac{\partial \phi}{\partial b_{1h}} N_{10} \right)^2 & \sum_{h=1}^{t} \left(\frac{\partial \phi^2}{\partial b_{1h}} N_{10} N_{11} x_{1h} \right) \\ \sum_{h=1}^{t} \left(\frac{\partial \phi^2}{\partial b_{1h}} N_{10} N_{11} x_{1h} \right) & \sum_{h=1}^{t} \left(\frac{\partial \phi}{\partial b_{1h}} N_{11} x_{1h} \right)^2 \\ \vdots & \cdots \\ \sum_{h=1}^{t} \left(\frac{\partial \phi}{\partial b_{1h}} \frac{\partial \phi}{\partial b_{mh}} N_{10} N_{mn} x_{nh} \right) & \sum_{h=1}^{t} \left(\frac{\partial \phi}{\partial b_{1h}} \frac{\partial \phi}{\partial b_{mh}} N_{11} N_{mn} x_{1h} x_{nh} \right) \end{bmatrix}$$

$$\begin{bmatrix} \sum_{h=1}^{t} \left(\frac{\partial \phi}{\partial b_{1h}} N_{10} \right) \left(\frac{\partial \phi}{\partial b_{1h}} N_{12} x_{2h} + \phi(b_{1h}) \right) & \cdots & \sum_{h=1}^{t} \left(\frac{\partial \phi}{\partial b_{1h}} \frac{\partial \phi}{\partial b_{mh}} N_{10} N_{mn} x_{nh} \right) \\ \vdots & \cdots & \vdots \\ \sum_{h=1}^{t} \left(\frac{\partial \phi}{\partial b_{1h}} N_{12} x_{2h} + \phi(b_{1h}) \right)^2 & \ddots & \vdots \\ \cdots & \cdots & \sum_{h=1}^{t} \left(\frac{\partial \phi}{\partial b_{mh}} N_{mn} x_{nh} \right)^2 \end{bmatrix} \tag{60}$$

Since J^{PLP} and Ω^{PLP} are convex functions, the regularization based on the least-squares solution as presented in (52) does not loose any potential solution.

$$\begin{aligned} l_\lambda = \arg\min \; J_\lambda^{PLP} &= \lambda J^{PLP} + (1-\lambda)\Omega^{PLP} \\ &= \lambda (Al - \tilde{y})^T (Al - \tilde{y}) + (1-\lambda) l^T Q l. \end{aligned} \tag{61}$$

where the optimum l, (i.e., l_λ), can be calculated by making the derivative of (61) equal to zero. The derivative of (61) in relation to l can be calculated as:

$$\frac{dJ_\lambda^{PLP}}{dl} = \lambda(-2A^T\tilde{y} + 2A^T A l) + (1-\lambda)2Ql \tag{62}$$

In order to find l_λ, the previous relation should be made equal to zero, as given below:

$$\lambda(-2A^T\tilde{y} + 2A^T A l) + (1-\lambda)2Ql = 0$$

$$-2\lambda A^T\tilde{y} + 2\lambda A^T A l + (1-\lambda)2Ql = 0$$

$$[\lambda A^T A + (1-\lambda)Q]l = \lambda A^T \tilde{y}$$

$$l_\lambda = [\lambda A^T A + (1-\lambda)Q]^{-1} \lambda A^T \tilde{y}, \tag{63}$$

if the matrix $[\lambda A^T A + (1-\lambda)Q]$ is non-singular. The Pareto-Optimum set can be found by varying λ between zero and one. This work applied the golden section algorithm in the validation error criteria to define λ_*. The validation error for the given formulation is a convex function of the linear parameters l.

5.2 Generalized Singular Value Decomposition

Consider the following properties of the Generalized Singular Value Decomposition (GSVD) Hansen (1998):

$$GSVD = \begin{cases} A = U_A S_A V^T \\ D = U_D S_D V^T \\ S_A^T S_A + S_D^T S_D = I \end{cases}, \tag{64}$$

where U is a unitary matrix, i.e., $U^{-1} = U^T$, and $S_A = diag(s_{A1},\ldots,s_{A(n+1)m})$, $S_D = diag(s_{D1},\ldots,s_{D(n+1)m})$ such that $s_{A1} \geq \ldots \geq s_{A(n+1)m} \geq 0$ and $s_{D(n+1)m} \geq \ldots \geq s_{D1} \geq 0$. Applying (64) in (63) the following is obtained:

$$\begin{aligned} l_\lambda &= \left[\lambda A^T A + (1-\lambda)Q\right]^{-1} \lambda A^T \tilde{y} \\ &= \lambda \left[\lambda V S_A^2 V^T + (1-\lambda) V S_D^2 V^T\right]^{-1} V S_A U_A^T \tilde{y} \\ &= \lambda \left[V(\lambda S_A^2 + (1-\lambda) S_D^2) V^T\right]^{-1} V S_A U_A^T \tilde{y} \\ &= \lambda \left[(\lambda S_A^2 + (1-\lambda) S_D^2) V^T\right]^{-1} V^{-1} V S_A U_A^T \tilde{y} \\ &= \lambda (V^T)^{-1} \left[\lambda S_A^2 + (1-\lambda) S_D^2\right]^{-1} S_A U_A^T \tilde{y} \end{aligned} \tag{65}$$

where $\left[\lambda S_A^2 + (1-\lambda) S_D^2\right]$ is a diagonal matrix with elements $\left[\lambda s_{Ai}^2 + (1-\lambda) s_{Di}^2\right]$. The unfiltered solution, disregarding the complexity control, i.e., $\lambda = 1$, is equal to

$$l_*(\lambda = 1) = (V^T)^{-1} S_A^{-1} U_A^T \tilde{y}. \tag{66}$$

The Wiener filter weights are evaluated comparing the unfiltered solution with the general solution of (65). The following is obtained

$$\begin{aligned} \Psi_i &= \frac{\lambda s_{Ai}^2}{\lambda s_{Ai}^2 + (1-\lambda) s_{Di}^2} \\ &= \frac{1}{1 + \lambda' \frac{s_{Di}^2}{s_{Ai}^2}} \end{aligned}, \tag{67}$$

where $\lambda' = (1-\lambda)/\lambda$, $\lambda \neq 0$ and s_{A_i}/s_{Di} are the generalized singular values. Similarly to the results using the simple SVD, the components with smaller singular values are filtered the most. Differently from the traditional Wiener filter, which only considers $s_{Di} = 1$, the MGM approach computes a general s_{Di}. It is possible to obtain $s_{Di} = 1$ using a identity matrix in the Q-norm. The Wiener filter weights define the relevance of each nonlinear neuron, filtering the unnecessary ones.

6. Results for benchmark problems

This section presents some experimental results in benchmarking problems considering the proposed ideas. Sigmoidal logistic functions have been used as PLP nonlinear activation function. Data sets from Intelligent data Analysis (IDA) repository are considered here as

presented in K. Muller & Scholkopf (2002). Table 1 summarizes the dimensionality of the input space, the number of training and test samples and the number of realizations for each data set. The results obtained by the PLP-MGM are compared with the results obtained by using the following machine learning techniques: (i) Support Vector Machine (SVM), (ii) kernel Fisher Discriminant (KFD), and (iii) Regularized AdaBoost (ABR) extracted from Muller et al. (2001); (iv)Leave-One-Out KFD (LKFD), and (v) Single objective Parallel Layer Perceptron (PLP) from Caminhas et al. (2003). The results are presented in Table 2.

Name	Dimension	Train	Test	Realizations
Banana	2	400	4900	100
B.Cancer	9	200	77	100
Diabetes	8	468	300	100
German	20	700	300	100
Heart	13	170	100	100
Image	1300	1010	18	20
S. Flare	9	666	400	100
Thyroid	5	140	75	100
Titanic	3	150	2051	100
Twonorm	20	400	700	100

Table 1. Ida repository data set summary.

	SVM	KFD	ABR	LKFD	PLP	PLP-MGM
Banana	11.5±0.7	10.8±0.5	10.9±0.4	10.4±0.4	10.7±.06	10.7±0.6
B. Cancer	26±5	26±5	27 ± 5	26 ± 4	27 ± 5	25 ± 4
Diabetes	23±2	23±2	24±2	23±2	23±2	23±2
German	24±2	24±2	24±2	24±2	30±3	24±2
Heart	16±3	16±4	17±4	16±4	19±3	16±3
Image	3.0±0.6	3.3±0.6	2.7±0.6	4.0±0.6	5±4	3.3±0.7
S. Flare	32±2	33±2	34±2	34±2	37±2	33±2
Thyr.	5±2	4±2	5±2	5±2	4±2	4±2
Titanic	22±1	23±2	23±1	22±1	23±1	22±1
Twon.	3.0±0.2	2.6±0.2	2.7±0.2	2.7±0.2	2.8±0.3	2.6±0.3

Table 2. Ida repository results.

The first noticeable result of Table 2 is that the PLP-MGM has outperformed the conventional PLP in most of the tested examples, and that PLP has never outperformed PLP-MGM. It is clear as well that the PLP-MGM has achieved similar results compared to those produced by the other approaches used for comparison.

7. Denoising Ground Penetrating Radar data

This section considers denoising Ground Penetrating Radar (GPR) using the PLP-MGM technique. This noise can be due to environmental conditions, geometric variations, and sensors characteristics. The numerical simulation follows the results described in Travassos et al. (2008). A block diagram of a typical GPR system to detect underground targets, is given in Fig. 3.

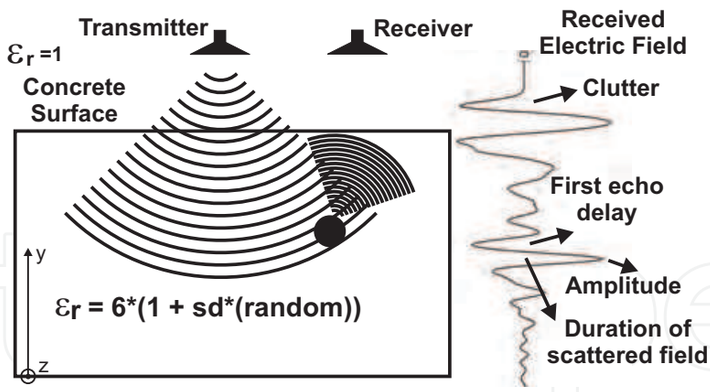

Fig. 3. The GPR problem.

The proposed configuration is tested to filter the noise of the scattered wave from a cylindrical air inclusion buried in a non-homogenous host medium, Vieira et al. (2009). Tables 3 and 4 considers white and colored Gaussian noise respectively. As the noise is stochastic by nature a statistical evaluation of the results is necessary. The simulations were done considering 20 different noises for each SNR, and a Neural Network trained for each of them. The results are presented in 3 and 4 they show a considerable improvement in the SNR, showing the effectiveness of the proposed approach.

	SNR in the Filtered Wave (dB)		
Noise (dB)	Mean	Max	Min
3	14.16	14.65	13.47
6	14.69	15.07	14.14
9	16.55	17.67	15.65
10	20.47	21.35	18.67

Table 3. SNR considering the GPR processed wave (filtered) by the proposed approach corrupted by White Gaussian Noise.

	SNR in the Filtered Wave (dB)		
Noise (dB)	Mean	Max	Min
3	12.76	13.22	11.96
6	15.30	16.21	14.17
9	20.36	20.73	19.96
10	20.58	20.93	20.09

Table 4. SNR considering the GPR wave processed (filtered) by the proposed approach corrupted by Colored Gaussian Noise.

8. Final Comments

This chapter described the use of the multiobjective optimization framework to train the Parallel Layer Perceptron network. This is based on the general concept that learning depends on two functions: the empirical risk and the network complexity. A formulation based on

the Tikhonov's regularization was proposed using a Q-norm as a complexity measure. This has a least-squares like closed form solution; therefore, it relies on simple computational algorithms. Moreover, it bores the good aspects of the Tikhonov's method. It opens discussions about other possible definitions of the matrix Q, different configurations of the PLP layers among others.

The results presented proved the effectiveness of the proposed approach. A wide comparison considering several benchmarking problems and algorithms were presented. Also a complex engineering problem was successfully solved using the proposed approach.

The relationships between the classical regularization, the structural risk minimization and the multiobjective formulation were also explored. These help the understanding concerning the nature of learning and their possibilities. It shows that the convexity is an important issue to the use of the WD method to MLPs. This is indeed a wide subject, and, due to space constraints, this chapter discussed a rather biased point-of-view on those subjects.

9. Acknowledgment

This work was supported by CNPq and FAPEMIG, Brazil.

10. References

Alavetti, D. C. & Eichel, L. R. (2004). Tikhonov regularization with a solution constraint, *SIAM J. Sci. Comput.* (26).

Bartlett, P. L. (1998). The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network., *IEEE Trans. on Information Theory* **2**(44): 525–536.

Caminhas, W. M., Vieira, D. A. G. & Vasconcelos, J. A. (2003). Parallel layer perceptron, *Neurocomputing* **3-4**(55): 771– 778.

Hansen, P. C. (1998). *Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA. ISBN 0-89871-403-6.

Hastie, T., Tibshirani, R. & Friedman, J. H. (2001). *The Elements of Statistical Learning*, first edn, Springer.

Hinton, G. E. (1989). Connections learning procedures, *Artificial inteligence* **40**: 185–234.

Ivanov, V. V. (1962). On linear problems which are not well-posed, *Soviet Math. Docl.* **3**(4): 981–983.

Ivanov, V. V. (1976). *The theory of approximate methods and their application to the numerical solution of singular integral equations*, Leyden : Noordhoff International. ISBN: 9028600361.

K. Muller, S. Mika, G. R. K. T. & Scholkopf, B. (2002). Ida bechmark repository used in several boosting, kfd and svm papers, *Technical report*. `ida.first.gmd.de/~raetsch/data/benchmarks.htm`.

Muller, K., Mika, S., Ratsh, G., Tsuda, K. & Scholkopf, B. (2001). An introduction to kernel-based learning algorithms, *IEEE Trans. on Neural Networks* **12**(2): 181–201.

Phillips, D. Z. (1962). A technique for numerical solution of certain integral equation of the first kind, *J. Assoc. Comput. Mach* **9**: 84–96.

S. Geman, E. B. & Doursat, R. (1992). Neural networks and the bias-variance dilemma, *Neural Computation* **1**(4): 1–58.

Shawe-Taylor, J. & Bartlett, P. L. (1998). Structural risk minimization over data-dependent hierarchies, *IEEE Trans. on Information Theory* **44**(5): 1926–1940.

Tikhonov, A. N. (1963). On solving ill-posed problem and the method of regularization, *Doklady Akademii Nauk USSR* **153**: 501–504.

Tikhonov, A. N. & Arsenin, V. Y. (1977). *Solution of ill-posed problems*, W. H. Winston, Washington, DC.

Travassos, X., Vieira, D., Ida, N., Vollaire, C. & Nicolas, A. (2008). Characterization of inclusions in a nonhomogeneous gpr problem by artificial neural networks, *Magnetics, IEEE Transactions on* **44**(6): 1630–1633.

Vapnik, V. N. (1992). Principles of structural risk minimization for learning theroy, *Advances in Neural Information Processing Systems* **4**: 831–838.

Vapnik, V. N. (1998). *Statistical Learning Theory*, New York: Wiley.

Vapnik, V. N. (2001). *The Nature of Statistical Learning Theory (Statistics for Engineering and Information Science)*, second edn, Springer.

Vasin, V. V. (1970). Relationship of several varitional methods for approximate solutions of ill-posed problems, *Math Notes* **7**: 161–166.

Vieira, D. A. G., Takahashi, R. H. C., Palade, V., Vasconcelos, J. A. & Caminhas, W. M. (2008). The *Q*-norm complexity measure and the minimum gradient method: A novel approach to the machine learning structural risk minimization problem, *Neural Networks, IEEE Transactions on* **19**(8): 1415–1430.

Vieira, D., Travassos, L., Saldanha, R. & Palade, V. (2009). Signal denoising in engineering problems through the minimum gradient method, *Neurocomputing* **72**(10-12): 2270 – 2275.

Machine Learning

Edited by Yagang Zhang

ISBN 978-953-307-033-9
Hard cover, 438 pages
Publisher InTech
Published online 01, February, 2010
Published in print edition February, 2010

Machine learning techniques have the potential of alleviating the complexity of knowledge acquisition. This book presents today's state and development tendencies of machine learning. It is a multi-author book. Taking into account the large amount of knowledge about machine learning and practice presented in the book, it is divided into three major parts: Introduction, Machine Learning Theory and Applications. Part I focuses on the introduction to machine learning. The author also attempts to promote a new design of thinking machines and development philosophy. Considering the growing complexity and serious difficulties of information processing in machine learning, in Part II of the book, the theoretical foundations of machine learning are considered, and they mainly include self-organizing maps (SOMs), clustering, artificial neural networks, nonlinear control, fuzzy system and knowledge-based system (KBS). Part III contains selected applications of various machine learning approaches, from flight delays, network intrusion, immune system, ship design to CT and RNA target prediction. The book will be of interest to industrial engineers and scientists as well as academics who wish to pursue machine learning. The book is intended for both graduate and postgraduate students in fields such as computer science, cybernetics, system sciences, engineering, statistics, and social sciences, and as a reference for software professionals and practitioners.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen