Kernel CMAC: an Efficient Neural Network for Classification and Regression

Gábor Horváth

Department of Measurement and Information Systems
Budapest University of Technology and Economics
Magyar tudósok körútja 2, H-1521 Budapest, Hungary
e-mail: horvath@mit.bme.hu

Abstract: Kernel methods in learning machines have been developed in the last decade as new techniques for solving classification and regression problems. Kernel methods have many advantageous properties regarding their learning and generalization capabilities, but for getting the solution usually the computationally complex quadratic programming is required. To reduce computational complexity a lot of different versions have been developed. These versions apply different kernel functions, utilize the training data in different ways or apply different criterion functions. This paper deals with a special kernel network, which is based on the CMAC neural network. Cerebellar Model Articulation Controller (CMAC) has some attractive features: fast learning capability and the possibility of efficient digital hardware implementation. Besides these attractive features the modelling and generalization capabilities of a CMAC may be rather limited. The paper shows that kernel CMAC – an extended version of the classical CMAC network implemented in a kernel form – improves that properties of the classical version significantly. Both the modelling and the generalization capabilities are improved while the limited computational complexity is maintained. The paper shows the architecture of this network and presents the relation between the classical CMAC and the kernel networks. The operation of the proposed architecture is illustrated using some common benchmark problems.

Keywords: kernel networks, input-output system modelling, neural networks, CMAC, generalization error

1 Introduction

Kernel machines like Support Vector Machines (SVMs) [1], Least Squares SVMs (LS-SVMs) [2] and the method of ridge regression [3] have beed developed in the last decade and proved to be efficient new approaches for solving the learning problem from samples. Kernel machines can be applied for linear and nonlinear classification and function approximation, so they can be used for solving problems that can be solved successfully with classical neural networks too. The basic idea of kernel machines is that they apply two consequtive mappings. The first one maps the points of the input space (the input data) into an intermediate space called feature space. The goal of this mapping is to transform the original nonlinear problem into a linear one. More exactly, if a problem in the original representation can only be solved by nonlinear approaches, its transformed version in the feature space can be solved using linear methods (classification or regression). The theoretical background of applying nonlinear mapping to transform a nonlinear problem into a linear one goes back to Cover's theorem on the separability of patterns [4]. Based on this theorem it can be stated that it is more likely to represent a classification problem in a linearly separable way in a high-dimensional space than in a low-dimensional one.

Many neural network architectures apply this idea when first the input vectors are transformed into a higher-dimensional feature space, then in the feature space a linear solution is derived. These networks has two layers: the first one is responsible for the nonlinear dimension-increasing mapping and the second is a simple layer composed of linear neurons. Such networks are the popular radial basis function (RBF) networks, and all other basis function networks, but Cerebellar Model Articulation Controller (CMAC) network can also be interpreted in this way. The drawback of this approach is that in many cases the dimension of the feature space may be extremely large – even infinite. So to look for a solution in the high-dimensional feature space is impractical or in many cases it is practically impossible.

Kernel machines solve the dimensionality-problem by applying a trick, which is called kernel trick. They also apply nonlinear mapping from input space into feature space, however, they do not look for the solution in the feature space, instead the solution is obtained in the kernel space, which is defined easily. The significance of using the kernel trick is that the complexity of the solution is greatly reduced: the dimension of the kernel space is upper bounded by the number of training samples independently of the dimension of the feature space. The mappings of a kernel machine for a classification problem is shown in Fig. 1.

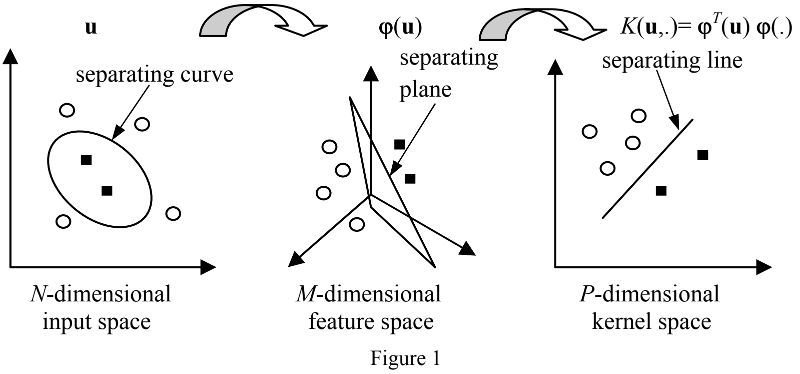

u
φ(**u**)
$K(\mathbf{u},.)=\varphi^T(\mathbf{u})\,\varphi(.)$
separating curve
separating plane
separating line
N-dimensional input space
M-dimensional feature space
P-dimensional kernel space

Figure 1
The mappings of a kernel machine

As it can be seen, in the input space only nonlinear separation is possible. In the feature space – where usually $M > N$ – the transformed points are linearly separable. The linear separability is maintained in the kernel space too, while the dimension of the kernel space is upper bounded by P, the number of available data points.

Cerebellar Model Articulation Controller (CMAC) [5] – a special feed-forward neural architecture, which belongs to the family of feed-forward networks with a single linear trainable layer – has some attractive features. The most important ones are its extremely fast learning capability and the special architecture that lets effective digital hardware implementation possible [6]. The CMAC architecture was proposed by Albus in the middle of the seventies [5] and it is considered as a real alternative to MLP and other feed-forward neural networks [7]. Although the properties of a CMAC were analysed mainly in the nineties (see eg. [8]-[11]), some interesting features were only recognized in the recent years. These results show that the attractive properties of the CMAC have a price: its modelling capability is inferior to that of an MLP. This is especially true for multivariate cases, as multivariate CMACs can learn to reproduce the training points exactly only if the training data come from a function belonging to the additive function set [8].

The modelling capability can be improved if the complexity of the network is increased. This more complex network was proposed in [9], but as the complexity of the CMAC depends on the dimension of the input data, in multivariate cases the high complexity can be an obstacle of implementation in any way. A further deficiency of CMAC is that its generalization capability is also inferior to that of an MLP even for univariate cases. The real reason of this property was shortly presented in [11] and a modified training algorithm was proposed for improving the generalization capability. This training algorithm is derived using a regularized [12] loss function, where the regularization term has some weight-smoothing effect.

This paper presents a different interpretation of the CMAC networks and details why this interpretation can help to improve the quality of the network without increasing the complexity even in multidimensional cases. The paper shows that this new interpretation corresponds to a kernel machine with second order B-spline kernel functions. The kernel interpretation may suffer from the same poor generalization capability, however the weight-smoothing regularization can be applied for the kernel CMAC too. This means that using kernel CMAC both the modelling and the generalization capabilities can be improved significantly. Moreover it can be shown that similarly to the original CMAC the kernel versions can also be trained iteratively, which may be important in such applications where real-time on-line adaptation is required.

The paper is organized as it follows. Section 2 summarizes some important features of kernel machines. Section 3 gives a short introduction to CMAC networks. This section presents the drawbacks of the classical CMAC. Section 4 shows how a CMAC can be interpreted as a kernel machine and it derives the main results related to the regularized version. Section 5 gives some illustrative examples to show how the proposed network can improve the capability of the network.

2 Kernel Machines

The goal of a kernel machine is to approximate a (nonlinear) function $y_d = f(\mathbf{u})$ ($\mathbf{u} \in \Re^N$, $y_d \in \Re$) using a training data set $\{\mathbf{u}(k), y_d(k)\}_{k=1}^{P}$. A kernel machine can be used to solve classification or regression problems. For classification the function to be approximated is $f \ : \ \Re^N \to \{\pm 1\}$, while for regression problems a continuous function $f \ : \ \Re^N \to \Re$ should be approximated. In the kernel machines first the **u** input vectors are projected into a higher dimensional feature space, using a set of nonlinear functions $\boldsymbol{\varphi}(\mathbf{u}) \ : \ \Re^N \to \Re^M$, then the output is obtained as a linear combination of the projected vectors [1]:

$$y(\mathbf{u}) = \sum_{j=1}^{M} w_j \varphi_j(\mathbf{u}) + b = \mathbf{w}^T \boldsymbol{\varphi}(\mathbf{u}) + b \tag{1}$$

where **w** is the weight vector and *b* is a bias term. The dimensionality (*M*) of the feature space is not defined directly, it follows from the method (it can even be infinite). The kernel trick makes it possible to obtain the solution not in the feature space but in the kernel space

$$y(\mathbf{u}) = \sum_{k=1}^{P} \alpha_k K(\mathbf{u}, \mathbf{u}(k)) + b \tag{2}$$

where the kernel function is formed as

$$K(\mathbf{u}(k), \mathbf{u}(j)) = \boldsymbol{\varphi}^T(\mathbf{u}(k)) \boldsymbol{\varphi}(\mathbf{u}(j)) \tag{3}$$

In (2) the α_k coefficients serve as the weight values in the kernel space. The number of these coefficients equals to or less (in some cases it may be much less) then the number of training points [1]. The main advantage of a kernel machine is that the kernel function can be defined directly without using the feature space representation. For this purpose the kernel function should fulfill some conditions [13]. Kernel machines can be constructed using constrained optimization, where first a criterion function and some constrainst are defined, and where the solution is obtained using a Lagrange multiplicator approach.

Kernel machines have many different versions. These versions apply different kernel functions or formulate the constrained optimization problem in different ways. Most often Gaussian kernels are used, but polynomial, spline, etc. kernels can also be applied [13]. The complexity of the solution depends on the form of the constraints. Using inequality constraints quadratic programming is required to reach the solution [1]. This approach was introduced by Boser, Guyon and Vapnik [14] and it results in the classical support vector machine (SVM) solution. A less complex solution is obtained if instead of the inequality constraints equality ones are applied. One approach of this version is when quadratic criterion function and equality constraints are used. It is called least squares support vector machine (LS-SVM) [2]. Ridge regression is similar to LS-SVM, although its derivation is slightly different from that of the LS-SVM [3]. Both in ridge regression and in LS-SVM instead of quadratic programming the solution can be obtained using simple matrix inversion. To show the detailes of the kernel machines is beyond the scope of this paper. These detailes can be found in the recently published excellent books and papers. See e.g. [13], [15], [16], [17].

3 A Short Overview of the CMAC

CMAC is an associative memory type neural network, which performs two subsequent mappings. The first one – which is a non-linear mapping – projects an input space point $\mathbf{u}\in\Re^N$ into an association vector **a**. The second mapping calculates the output $y\in\Re$ of the network as a scalar product of the association vector **a** and the weight vector **w**:

$$y(\mathbf{u})=\mathbf{a}(\mathbf{u})^{\mathrm{T}}\mathbf{w} \tag{4}$$

The association vectors are sparse binary vectors, which have only *C* active elements: *C* bits of the association vector are ones and the others are zeros. As the association vectors are binary ones, scalar products can be implemented without any multiplication; the scalar product is nothing more than the sum of the weights selected by the active bits of the association vector.

$$y(\mathbf{u})=\sum_{i:a_i(\mathbf{u})=1} w_i \tag{5}$$

CMAC uses quantized inputs, so the number of the possible different input data is finite. There is a one-to-one mapping between the discrete input data and the association vectors, i.e. each possible input point has a unique association vector representation.

Another interpretation can also be given to the CMAC. In this interpretation for an *N*-variate CMAC every bit in the association vector corresponds to a binary basis function with a compact N-dimensional hypercube support. The size of the hypercube is C quantization intervals. This means that a bit will be active if and only if the input value is within the support of the corresponding basis function. This support is often called receptive field of the basis function [5].

The mapping from the input space into the association vector should have the following characteristics:

(i) it should map two neighbouring input points into such association vectors that only a few elements – i.e. few bits – are different,

(ii) as the distance between two input points grows, the number of the common active bits in the corresponding association vectors decreases. For input points far enough from each other – further then the neighbourhood determined by the parameter C – the association vectors should not have any common bits.

This mapping is responsible for the non-linear property and the generalization of the whole system. The first layer implements a special encoding of the quantized input data. This layer is fixed. The trainable elements, the weight values that can be updated using the simple LMS rule, are in the second layer. The way of encoding, the positions of the basis functions in the first layer, and the value of C determine the generalization property of the network. In one-dimensional cases every quantization interval will determine a basis function, so the number of basis functions is approximately equal to the number of possible discrete inputs. However, if we follow this rule in multivariate cases – the resulted network will be called full-overlay CMAC – , the number of basis functions will grow exponentially with the number of input variables, so the network may become too complex. As every selected basis function will be multiplied by a weight value, the size of the weight memory is equal to the total number of basis functions, to the length of the association vector. If there are r_i discrete values for the i-th input dimension, an N-dimensional CMAC needs $M = \prod_{i=1}^{N}(r_i + C - 1)$ weight values. In multivariate cases the weight memory can be so huge that practically it cannot be implemented.

To avoid this high complexity the number of basis functions must be reduced. In a classical multivariate CMAC this reduction is achieved by using basis functions positioned only at the diagonals of the quantized input space. The positions of the overlays and the basis functions of one overlay can be represented by definite points. In the original Albus scheme the overlay-representing points are in the main diagonal of the input space, while the basis-function-positions are represented by the sub-diagonal points (see Fig. 2).

The shaded regions in Fig. 2 are the receptive fields of different basis functions. As it is shown the basis functions are grouped into overlays. One overlay contains basis functions with non-overlapping supports, but the union of the supports covers the whole input space. The different overlays have the same structure; they consist of similar basis functions in shifted positions. Every input data will select C basis functions, each of them on a different overlay, so in an overlay one and only one basis function will be active for every input point.

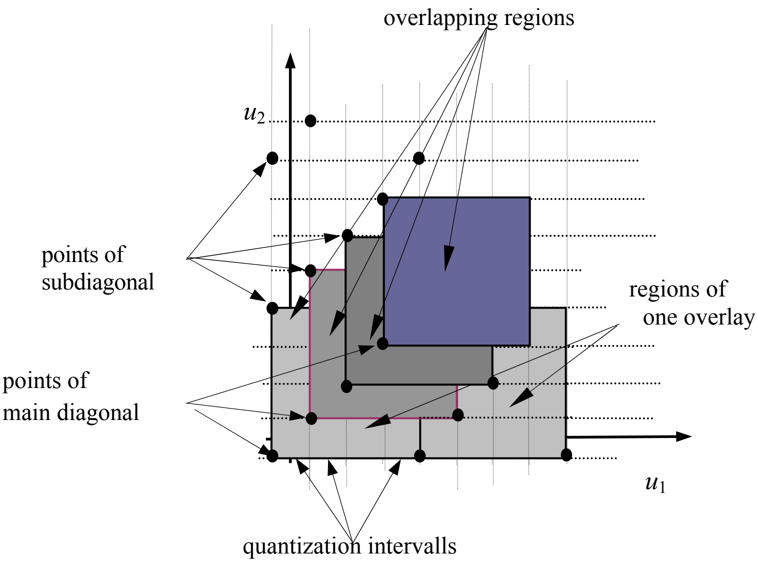


Figure 2
The basis functions of a two variable CMAC

In the original Albus architecture the number of overlays does not depend on the dimension of the input vectors; it is always C. This means that in multivariate cases the number of basis function will not grow exponentially with the input dimension, it will be "only" $M = \left\lceil \frac{1}{C^{N-1}} \prod_{i=1}^{N} (r_i + C - 1) \right\rceil$. This is an advantageous property from the point of view of implementation, however this reduced number of basis functions is the real reason of the inferior modelling capability of the multivariable CMACs, as reducing the number of basis functions the number of free parameters will also be reduced. Here modelling capability refers to the ability that a network can learn to reproduce exactly the training data: a network with this ability will have no modelling error.

The consequence of the reduced number of basis functions is that an arbitrary classical binary multivariate CMAC can reproduce exactly the training points only if they are obtained from an additive function [8]. For more general cases there will be modelling error i.e. error at the training points. It should be mentioned that in multivariate cases even this reduced weight memory may be too large, so further complexity reduction may be required. As an example consider a tendimensional binary CMAC where all input components are quantized into 10 bits. In this case the length of the association vector would be approximately 2^{55}. Although this is a greatly reduced value compared to 2^{100}, which is the weight memory size of the full-overlay version, this value is still prohibitively large.

Further reduction is achieved by applying a new compressing layer [4], which uses hash-coding. Although hash-coding solves the complexity problem, it can result in collisions of the mapped weights, and some unfavourable effects on the convergence of CMAC learning [18], [19]. As it will be seen later the proposed new interpretation solves the complexity problem without the application of hash-coding, so we will not deal with this effect.

Another way of avoiding the complexity problem is to decompose a multivariate problem into many one-dimensional ones, so instead of implementing a multidimensional CMAC it is better to implement many simple one-dimensional networks. The resulted hierarchical, tree-structured network – called MS_CMAC [20] – can be trained using time inversion technique [21]. MS_CMAC greatly reduces the complexity of the network, however there are some restrictions in its application as it can be applied only if the training points are positioned at regular grid-points. A further drawback is that most of the simple networks need training even in the recall phase, increasing the recall time significantly.

A CMAC – as it has a linear output layer – can be trained by the LMS algorithm:

$$\mathbf{w}(k+1) = \mathbf{w}(k) + \mu\, \mathbf{a}(k)e(k), \tag{6}$$

where $e(k) = y_d(k) - y(k) = y_d(k) - \mathbf{w}^T\mathbf{a}(k)$ is the error at the k-th training step. Here $y_d(k)$ is the desired output for the k-th training point, $y(k)$ is the network output for the same input, $\mathbf{a}(k) = \mathbf{a}(\mathbf{u}(\mathrm{k}))$ and μ is the learning rate. Training will minimize the quadratic error

$$\min_{\mathbf{w}} J = \frac{1}{2}\sum_{k=1}^{P} e(k)^2 \tag{7}$$

where P is the number of training points.

The solution of the training can also be written in a closed form

$$\mathbf{w}^* = \mathbf{A}^\dagger \mathbf{y}_d \tag{8}$$

where $\mathbf{A}^\dagger = \mathbf{A}^T\left(\mathbf{A}\mathbf{A}^T\right)^{-1}$ is the pseudo inverse of the association matrix formed from the association vectors $\mathbf{a}(i) = \mathbf{a}(\mathbf{u}(i))$, and $\mathbf{y}_d{}^T = [y_d(1)\ \ y_d(2)\ \ \ldots\ \ y_d(P)]$ is the output vector formed from the desired values of all training data. The response of the trained network for a given input $\mathbf{u}$ can be determined using the solution weight vector:

$$y(\mathbf{u}) = \mathbf{a}^T(\mathbf{u})\mathbf{w}^* = \mathbf{a}^T(\mathbf{u})\mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{y}_d \quad . \tag{9}$$

4 Kernel CMAC

4.1 The Derivation of the Kernel CMAC

The relation between CMACs and kernel machines can be shown if we recognize that the association vector of a CMAC corresponds to the feature space representation of the kernel machines. This means that the non-linear functions that map the input data points into the feature space are the rectangular basis functions. The binary basis functions can be regarded as first-order B-spline functions of fixed positions.

To get the kernel representation of the CMAC we should apply (3) for the binary basis function. In univariate cases second-order B-spline kernels can be obtained where the centre parameters are the input training points. Fig. 3 shows a first-order B-spline basis function (a), and the corresponding kernel function – a second order B-spline – (b).

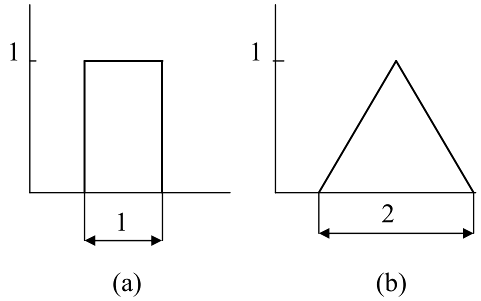
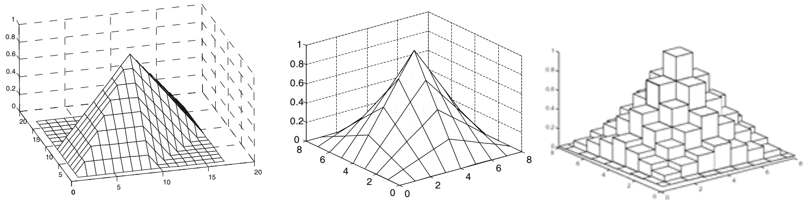


Figure 3
First-order (a) and second-order (b) B-spline function

In kernel representation the number of kernel functions does not depend on the number of basis functions in the feature space; because of the kernel trick it is always upper bounded by the number of training points *P.* This means that there are no reason to reduce the number of basis functions in the feature space: we can apply full-overlay CMAC. The only consequence is that the kernel function will be different and the modelling capability of the resulted network will be improved as a full-overlay CMAC can learn all training points exactly even in multivariate cases. Fig. 4 shows the 2D kernel function for classical CMAC (a), and for a full-overlay CMAC (b). This latter can be obtained as a tensor product of the two one-dimensional second order B-spline functions. Fig. 4(c) shows the quantized version of the 2D full-overlay kernel function. As in a CMAC quantized input data are used, this function is used as a kernel function in the proposed kernel CMAC.

The kernel interpretation can be extended to higher-order CMACs too [9] where higher order basis functions (*k*-th order B-splines with support of *C*) are applied. In these cases CMACs correspond to kernel machines with properly chosen higher-order (2*k*-th order) B-spline kernels.

Kernel machines can be derived through constrained optimisation. The different versions of kernel machines apply different loss functions. Vapnik's SVM for regression applies ε-insensitive loss function [1], while LS-SVM can be obtained if quadratic loss function is used [2]. The classical CMAC uses quadratic loss function too, so we obtain an equivalent kernel representation if in the constrained optimisation also quadratic loss function is used. This means that the kernel CMAC can be considered as a special LS-SVM.



(a) (b) (c)

Figure 4

2D kernel functions classical CMAC (a), full-overlay CMAC (b), and its spatially quantized version (c)

The response of a trained network for a given input can be obtained by (9). To see that this form can be interpreted as a kernel solution do construct an LS-SVM network with similar feature space representation. For LS-SVM regression we seek for the solution of the following constrained optimisation.

$$\min_{\mathbf{w}} J\left(\mathbf{w},\mathbf{e}\right)=\frac{1}{2}\mathbf{w}^T\mathbf{w}+\frac{\gamma}{2}\sum_{k=1}^{P}e\left(k\right)^2 \tag{10}$$

such that $y_d(k)=\mathbf{w}^T\mathbf{a}(k)+e(k)$. Here there is no bias term, as in the classical CMAC bias term is not used. The problem in this form can be solved by constructing the Lagrangian

$$L(\mathbf{w},\mathbf{e},\alpha)=J(\mathbf{w},\mathbf{e})-\sum_{k=1}^{P}\alpha_k\left(\mathbf{w}^T\mathbf{a}(k)+e(k)-y_d(k)\right) \tag{11}$$

where α_k are the Lagrange multipliers. The conditions for optimality can be given by

$$\begin{cases} \dfrac{\partial L(\mathbf{w},\mathbf{e},\alpha)}{\partial \mathbf{w}} = \mathbf{0} \;\rightarrow\; \mathbf{w} = \sum_{k=1}^{P} \alpha_k \mathbf{a}(k) \\ \dfrac{\partial L(\mathbf{w},\mathbf{e},\alpha)}{\partial e(k)} = 0 \;\rightarrow\; \alpha_k = \gamma e(k) \qquad\qquad k = 1,...,P \\ \dfrac{\partial L(\mathbf{w},\mathbf{e},\alpha)}{\partial \alpha_k} = 0 \rightarrow\; \mathbf{w}^T \mathbf{a}(\mathbf{u}(k)) + e(k) - y_d(k) = 0 \quad k = 1,...,P \end{cases} \tag{12}$$

Using the results of (12) in (11) the Lagrange multipliers can be obtained as a solution of the following linear system

$$\left[\mathbf{K} + \frac{1}{\gamma}\mathbf{I}\right]\boldsymbol{\alpha} = \mathbf{y}_d \tag{13}$$

Here $\mathbf{K} = \mathbf{A}\mathbf{A}^T$ is the kernel matrix and **I** is a $P \times P$ identity matrix. The response of the network can be obtained as

$$\begin{aligned} y(\mathbf{u}) &= \mathbf{a}^T(\mathbf{u})\mathbf{w} = \mathbf{a}^T(\mathbf{u})\sum_{k=1}^{P} \alpha_k \mathbf{a}(k) = \sum_{i=1}^{P} \alpha_k K(\mathbf{u},\mathbf{u}(k)) = \mathbf{K}^T(\mathbf{u})\boldsymbol{\alpha} \\ &= \mathbf{a}^T(\mathbf{u})\mathbf{A}^T\left[\mathbf{K} + \frac{1}{\gamma}\mathbf{I}\right]^{-1}\mathbf{y}_d = \mathbf{a}^T(\mathbf{u})\mathbf{A}^T\left[\mathbf{A}\mathbf{A}^T + \frac{1}{\gamma}\mathbf{I}\right]^{-1}\mathbf{y}_d \end{aligned} \tag{14}$$

The resulted kernel machine is an LS-SVM or more exactly a ridge regression solution [3], because of the lack of the bias term. Comparing (9) and (14), it can be seen that the only difference between the classical CMAC and the ridge regression solution is the term $(1/\gamma)\mathbf{I}$, which comes from the modified loss function of (10). However, if the matrix $\mathbf{A}\mathbf{A}^T$ is singular or it is near to singular that may cause numerical stability problems in the inverse calculation, a regularization term must be used: instead of computing $(\mathbf{A}\mathbf{A}^T)^{-1}$ the regularized inverse $(\mathbf{A}\mathbf{A}^T + \eta\mathbf{I})^{-1}$ is computed, where η is the regularization coefficient. In this case the two forms are equivalent.

4.2 Kernel CMAC with Weight-smoothing

This kernel representation improves the modelling property of the CMAC. As it corresponds to a full-overlay CMAC it can learn all training data exactly. However, the generalization capability is not improved. In the derivation of the kernel machines regularization and the Lagrange multiplier approach are applied. To get a CMAC with better generalization capability a further regularization term can be applied. Smoothing regularization can be obtained if a new term is added to the loss function of (10). The modified optimization problem can be formulated as follows:

$$\min_{\mathbf{w}} J(\mathbf{w},\mathbf{e}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{\gamma}{2}\sum_{k=1}^{P} e_k^2 + \frac{\lambda}{2}\sum_{k=1}^{P}\sum_{i}\left(\frac{y_{d_k}}{C} - w_k(i)\right)^2 \tag{15}$$

$w_k(i)$ is a weight value selected by the *i*th active bit of $\mathbf{a}_k$, so *i* runs through the indexes where $a_k(i)=1$. As the equality constraint is the same as in (10), we obtain the Lagrangian

$$L(\mathbf{w},\mathbf{e},\alpha) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{\gamma}{2}\sum_{k=1}^{P} e_k^2 + \frac{\lambda}{2}\sum_{k=1}^{P}\sum_{i}\left(\frac{y_{d_k}}{C} - w_k(i)\right)^2 - \sum_{k=1}^{P}\alpha_k\left(\mathbf{w}^T\mathbf{a}_k + e_k - y_{d_k}\right) \tag{16}$$

The Lagrange multipliers can be obtained again as a solution of a linear system.

$$\boldsymbol{\alpha} = \left(\mathbf{K_D} + \frac{1}{\gamma}\mathbf{I}\right)^{-1}\left(\mathbf{I} - \frac{\lambda}{C}\mathbf{K_D}\right)\mathbf{y}_d \tag{17}$$

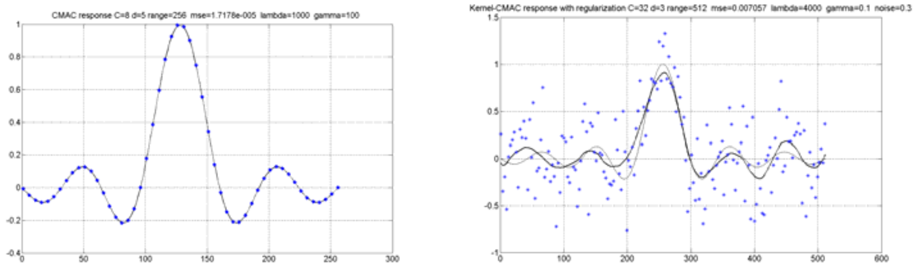
where $\mathbf{K_D} = \mathbf{A}\left(\mathbf{I}+\lambda\mathbf{D}\right)^{-1}\mathbf{A}^T$ and $\mathbf{D} = \sum_{k=1}^{P} diag(\mathbf{a}_k)$.

The response of the network becomes

$$y(\mathbf{u}) = \mathbf{a}^T(\mathbf{u})\left(\mathbf{I}+\lambda\mathbf{D}\right)^{-1}\mathbf{A}^T\left[\boldsymbol{\alpha} + \frac{\lambda}{C}\mathbf{y}_d\right]. \tag{18}$$

5 Illustrative Experimental Results

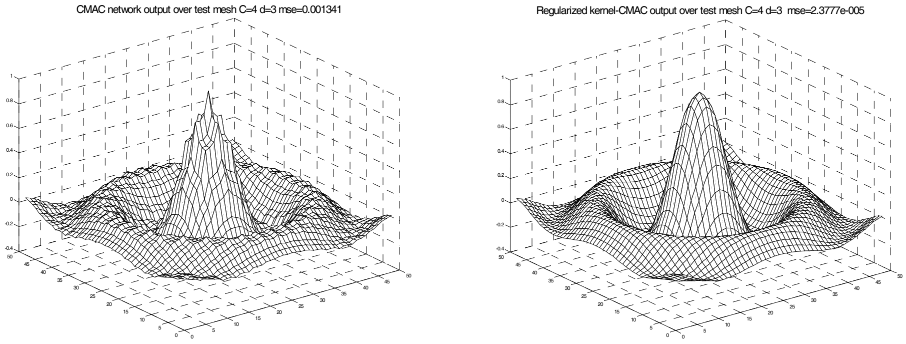
The different kernel versions of the CMAC network were validated by extensive experiments. Here only the results for some simple classification and regression benchmark problems are presented. The function approximation capability of the kernel CMAC is illustrated using the 1D (Fig. 5) and 2D (Fig. 6) *sinc* functions. For classification the two spiral problem (Fig. 7) is solved. This is a benchmark task, which is rather difficult for a classical MLP. These experiments show that the response of the reguralized kernel CMAC is much better than the response of the classical binary CMAC, the approximation or the classification error is significantly reduced.



(a) (b)

Figure 5

The response of the kernel CMAC with weight-smoothing regularization using noiseless (a), and noisy (b) training data. C =8, λ=10^3



(a) (b)

Figure 6

The response of the kernel CMAC without (a), and with weight-smoothing regularization. C =32, λ=10^3

Because of the finite support kernel functions local approximation and relatively low computational complexity are the additional advantages of kernel CMAC. Using this solution the large generalization error can be reduced significantly, so the regularized kernel CMAC is a real alternative of the popular neural network architectures like MLP and RBF even for multivariate cases.

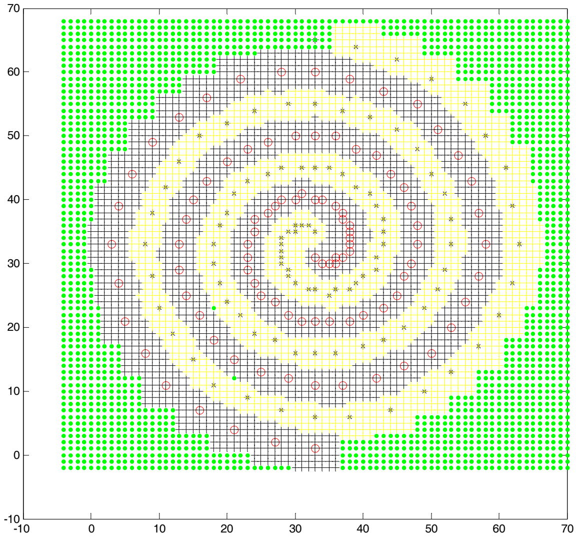


Figure 7

The solution of the two-spiral classification problem using kernel CMAC. *C*=4

Conclusions

In this paper it was shown that a CMAC network can be interpreted as a kernel machine with B-spline kernel function. This kernel interpretation makes it possible to increase the number of binary basis functions – the number of overlays, as in kernel interpretation the network complexity is upper bounded by the number of training samples, even if the number of binary basis functions of the original network is extremely large. The consequence of the increased number of basis function is that this version will have better modelling capability, and applying a special weight smoothing regularization the generalization capability can also be improved. Kernel CMAC can be applied successfully for both regression and classification problems even when high dimensional input vectors are used. The possibility of adaptive training ensures that the main advantages of the classical CMAC (adaptive operation, fast training, simple digital hardware implementation) can be maintained, although the multiplierless structure is lost.

References

[1] V. Vapnik: "Statistical Learning Theory", Wiley, New York, 1998

[2] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, B. and J. Vandewalle: "Least Squares Support Vector Machines", World Scientific, Singapore, 2002

[3] C. Saunders, A. Gammerman and V. Vovk: "Ridge Regression Learning Algorithm in Dual Variables. Machine Learning", *Proc. of the Fifteenth Int. Conf. on Machine Learning*, pp. 515-521, 1998

[4] T. M. Cover: "Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition" *IEEE Trans. on Electronic Computers*, EC-14, pp. 326-334, 1965

[5] J. S. Albus, "A New Approach to Manipulator Control: The Cerebellar Model Articulation Controller (CMAC)", *Transaction of the ASME*, pp. 220-227, Sept. 1975

[6] J. S. Ker, Y. H. Kuo, R. C. Wen and B. D. Liu: "Hardware Implementation of CMAC Neural Network with Reduced Storage Requirement", *IEEE Trans. on Neural Networks*, Vol. 8, pp. 1545-1556, 1997

[7] T. W. Miller III., F. H. Glanz and L. G. Kraft: "CMAC: An Associative Neural Network Alternative to Backpropagation" *Proceedings of the IEEE*, Vol. 78, pp. 1561-1567, 1990

[8] M. Brown, C. J. Harris and P. C. Parks, "The Interpolation Capabilities of the Binary Cmac", *Neural Networks,* Vol. 6, No. 3, pp. 429-440, 1993

[9] S. H. Lane, D. A. Handelman and J. J. Gelfand: "Theory and Development of Higher-Order CMAC Neural Networks", *IEEE Control Systems*, Vol. Apr., pp. 23-30, 1992

[10] C. T. Chiang and C. S. Lin: ''Learning Convergence of CMAC Technique'' *IEEE Trans. on Neural Networks, Vol. 8.* No. 6, pp. 1281-1292, 1996

[11] T. Szabó and G. Horváth: "Improving the Generalization Capability of the Binary CMAC" *Proc. Int. Joint Conf. on Neural Networks, IJCNN'2000,* Como, Italy, Vol. 3, pp. 85-90, 2000

[12] A. N. Tikhonov and V. Y. Arsenin: "Solution of Ill-posed Problems", Washington, DC, W. H. Winston, 1977

[13] B. Schölkopf and A. Smola: "Learning with Kernels. Support Vector Machines, Regularization, Optimization and Beyond" The MIT Press, Cambridbe, MA, 2002

[14] B. E. Boser, I. M. Guyon and V. N Vapnik: "A Training Algorithm for Optimal Margin Claasifiers" Fifth Annual Workhop on Computational Learning Theory, Pittsburg, ACM. pp. 144-152, 1992

[15] B. Schölkopf, C. J. C. Burges and A. J. Smola (eds.): "Advanced in Kernel Methods. Support Vector Learning" The MIT Press, Cambridbe, MA, 1999

[16] V. N. Vapnik: "The Nature of Statistical Learning Theory", Springer, 1995

[17] R. Herbrich: "Learning Kernel Classifiers, Theory and Algorithms", The MIT Pres, Cambridge, MA, USA, 2002

[18] L. Zhong, Z. Zhongming and Z. Chongguang: “The Unfavorable Effects of Hash Coding on CMAC Convergence and Compensatory Measure” *IEEE International Conference on Intelligent Processing Systems,* Beijing, China, pp. 419-422, 1997

[19] Z.-Q. Wang, J. L. Schiano and M. Ginsberg: “Hash Coding in CMAC Neural Networks” *Proc. of the IEEE International Conference on Neural Networks,* Washington, USA, Vol. 3, pp. 1698-1703, 1996

[20] J. C. Jan and S. L. Hung: High-Order MS_CMAC Neural Network, *IEEE Trans. on Neural Networks*, Vol. 12, No. 3, 2001, pp. 598-603

[21] J. S. Albus: "Data Storage in the Cerebellar Model Articulation Controller", *J. Dyn. Systems, Measurement Contr*. Vol. 97, No. 3, pp. 228-233, 1975

Dynamic System Using Conjunctive Operator

József Dombi

Árpád tér 2, H-6720 Szeged, Hungary, email: dombi@inf.u-szeged.hu

József D. Dombi

Árpád tér 2, H-6720 Szeged, Hungary, email: dombijd@inf.u-szeged.hu

Abstract: We present a tool to describe and simulate dynami systems. We use positive and negative influences. Our starting point is aggregation. We build positive and negative effects with proper transformations of the sigmoid function and using the conjunctive operator. From the input we calculate the output effect with the help of the aggregation operator. This algorithm is comparable with the concept of fuzzy cognitive maps.

Keywords: dynamic system, pliant concept, dombi operator

1 Introduction

We usually face serious difficulties when we handle sophisticated dynamic systems. Developing a model requires effort and specialized knowledge. Usually a system involves complicated causal chains, which might be nonlinear. It should also be mentioned, that numerical data may be hard to get, they can be even uncertain. Using a classical dynamic system model can be hard computationally. Fuzzy Cognitive Map (FCM) approach overcomes the above mentioned difficulties. FCM was proposed by Kosko [1][2][3] and is a hybrid method that lies in some sense between fuzzy systems and neural networks. Knowledge is represented in a symbolic manner using states, processes and events. All type of informations have numerical values. FCM allows us to perform qualitative simulations and experiment with a dynamic model. FCM has better properties than expert systems, since it is relatively easy to use them to represent structured knowledge and the inference can be computed by numeric matrix operation instead of applying rules. Our solution is similar to FCM. It is called Pliant Cognitive Map (PCM). It is based on a qualitative description of the influences, i.e. it is enough to know a rough description of the system. The main difference between PCM and FCM is the replacement of matrix multiplication by fuzzy operators.

2 Short Description of FCM

In FCM the causal relationship is expressed by either positive or negative signs ordered by different weights. As we mentioned this will be replaced by unary operators in PCM.

Let $\{C_1, . . . ,C_m\}$ be concepts. Define a directed graph over the concepts. We assign a weight $w_{ij} \in [0, 1]$ to the edge directed from concept C_i to concept C_j . The weight measures the influence of C_i on C_j . If the $w_{ij} = 1/2$ then w_{ij} is called neutral value. If $w_{ij} = 0$ then we say it is a maximum negative influence, if $w_{ij} = 1$ then we say it is a maximal positive influence or causality (In FCM $w_{ij} \in [-1,0, 1]$):

- $w_{ij} > 1/2$ indicates direct (positive) causality between concepts C_i and C_j. That is the increase (decrease) in the value of C_i leads to increase (decrease) on the value of C_j.
- $w_{ij} < 1/2$ indicates inverse(negative) causality between concepts C_i and C_j. That is the increase (decrease) in the value of C_i leads to decrease (increase) on the value of C_j.
- $w_{ij} = 1/2$ indicates C_i and C_j are neutral to each other.

In the pliant case w_{ij} depends on time (t), i. e.

$$w(t) = \left(w_{ij}(t)\right)_{n\times n} \tag{1}$$

The activation level a_i of concept C_i is calculated by an iteration process. In FCM

$$a_i^{(n)} = f\left(\sum_{i=1}^{n} w_{ij}\, a_i^{(n-1)}\right) \tag{2}$$

where a_i^n is the new activation level of concept C_i at time t+1, a_i^0 is the activation level of concept C_i at time t and f is the threshold function. FCM has the advantage that we obtain the new state vector by multiplying the previous state vector by the edge matrix W that shows the effect of the change in the activation level of one concept to another concept.

3 Basic Concept of PCM

In this paper we modify the concept of FCM. We use cognitive maps to represent knowledge and to model decision making, which was introduced by Axelrod. The cognitive map describes the whole system by a graph showing the cause effects along concepts. It is a directed graph with feedback, that describes the concepts of the world and the casual influences between the concepts. From logic point of view the causal concepts are unary operators of a continuous valued logic containing negation operators in the case of inhibition effects. The value of a node reflects the degree of the activity of the system at a particular time. Concept values are expressed on a normal [0, 1] range. Kosko used fuzzy values and matrix multiplication to calculate the next state of the systems.

We drop the concept of matrix multiplication. The matrix multiplication does not suit well in continuous logic (or fuzzy logic), where the truth value is 1 and the false is 0. General operators are more effective.

Logic and the cognitive map model correspond to each other in the PCM. It is easier to build up a PCM. After we identified the PCM with the real world, extracting the knowledge is easy. Combination of cognitive maps with logic helps us extracting knowledge more efficiently opposed to the use of rule based systems. The classical knowledge representation in expert systems is made through a decision tree. This form of knowledge presentation in most cases cannot model the dynamic behavior of the world. Instead of values, we use time dependent functions that are similar to impulse functions which represent positive and negative influences.

Values do not denote exact quantities, they denote the degree of activation. The inverse of normalization could express the values coming from the real world i.e. using sigmoid function. In spite of Fuzzy Cognitive Maps we do not use thresholds to force the values between 0 and 1. The mapping is a variation of the fuzzification process in fuzzy logic, furthermore it always destroys our desire to get quantitative results. In pliant logic we map the real world into logic. These maps are continuously strict monotonously increasing functions, and so the inverse of these functions yields the data of the real world.

In the pliant concept we aggregate the influences instead of summing up the values. The result always remains between 0 and 1, so we can avoid normalization as an additional step. The aggregation in pliant logic is a general operation, which contains conjunctive operators and disjunctive operators as well. Depending on the parameter called neutral value of the aggregation operator we build logical operators (Dombi operators).

Using PCM (Pliant Cognitive Maps) we answer what if questions based on an initial scenario. Let a_0 be the initial state vector. The new state is calculated repeatedly with the aggregation operator until the system converges i. e. $\left|a_i^0 - a_i^n\right| < \epsilon$.

We obtain the resulting equilibrium vector, which provides the answer to our what-if questions. The PCM can be used in all areas covered by FCM.

4 Components of PCM

4.1 Negation

The properties of negation are:

- defined on (0,1) and the values are also in (0,1)
- n(0) = 1
- n(1) = 0
- continuous
- strictly decreasing function
- involutive, n(n(x)) = x
- $n(\upsilon) = \upsilon_0$ or $n(\upsilon_*) = \upsilon_*$

The corresponding negation function in PCM is

$$n_\nu(x) = \frac{1}{1 + \frac{1-\nu_0}{\nu_0}\frac{1-\nu}{\nu}\frac{x}{1-x}} \tag{3}$$

$$n_{\nu_*}(x) = \frac{1}{1 + \left(\frac{1-\nu_*}{\nu_*}\right)^2 \frac{x}{1-x}} \tag{4}$$

4.2 Conjunction and Disjunction

Using an impulse function we can create positive and negative influences. Our task is to produce this function. The function is defined in time line and maps to [0, 1]. To construct this function we use a fuzzy operator and sigmoid function. It is important to consider that the fuzzy operator and sigmoid function fits well, accordingly we use Dombi[5] operator. The associative function equation is the following:

$$c(x,y) = f_c^{-1}(f_c(x) + f_c(y)) \quad d(x,y) = f_d^{-1}(f_d(x) + f_d(y)) \tag{5}$$

We call c(x,y) and d(x,y) pliant system if $f_c(x) + f_d(x) = 1$. The function of Dombi operators are pliant systems and:

$$f_c(x) = \frac{1-x}{x}, f_c^{-1}(x) = \frac{1}{1+x}, f_d(x) = \left(\frac{1-x}{x}\right)^{-1}, f_d^{-1}(x) = \frac{1}{1+x^{-1}} \tag{6}$$

If we use Dombi operator in the conjunction and disjunction function we get the following equation:

$$c(x,y) = \frac{1}{1+\frac{1-x}{x}+\frac{1-y}{y}},\ d(x,y) = \frac{1}{1+\left(\left(\frac{1-x}{x}\right)^{-1}+\left(\frac{1-y}{y}\right)^{-1}\right)^{-1}} \tag{7}$$

Using the Dombi operator we get the powered function in the following form:

$$o(x,y) = \frac{1}{1+\left(\left(\frac{1-x}{x}\right)^{\lambda}+\left(\frac{1-y}{y}\right)^{\lambda}\right)^{\frac{1}{\lambda}}} \tag{8}$$

If $\lambda > 0$ then it is conjunction function, if $\lambda < 0$ then it is disjunction function.

If we pay attention to the general form of the weight we can generalize the formula. In this case we lose the associativity property, however we get other good properties, which are very useful (i.e. idempotency). Let u and v be the weights, with the following qualities: u+v=1, 0<=u,v<=1. With these restriction we get the generalized function:

$$o_\lambda(u,x;v,y) = \frac{1}{1+\left(u\left(\frac{1-x}{x}\right)^{\lambda}+v\left(\frac{1-y}{y}\right)^{\lambda}\right)^{\frac{1}{\lambda}}} \tag{9}$$

4.3 Aggregation

Beside the developed logical operators in fuzzy theory, a non logical operator also appears. The reason for this is the insufficiency of using either conjunctive or disjunction operators for real world situations.

The aggregation operator is axiomatically based and it has several good properties as:

- defined on (0,1) and the values are also in (0,1)
- associativity
- strictly monotonously increasing
- continuous on [0,1) interval
- a(0,0) = 0 and a(1,1) = 1

The rational form of an aggregation operator [4] is:

$$a(x_1,\dots,x_n) = \frac{1}{1+\frac{1-\nu_0}{\nu_0}\left(\frac{\nu}{1-\nu}\right)^n \prod_{i=1}^n \frac{1-x_i}{x_i}} \quad (10)$$

Aggregation is connected to negation operators. The negation function and the aggregation operator are closely related. It can be easily seen that:

- n(a(x,y)) = a(n(x),n(y))
- a(x,n(x)) = v_0
- a(x,v_0) = x

These properties of the aggregation are natural:

- Aggregating positive values and negating is the same as aggregating negative values
- By aggregating a positive and its negated value we get the neutral values back
- Aggregating x with the neutral value we get back x.

We can model conjunctive and disjunctive operators with the aggregation operator. If *v* is close to 0 then the operation has a disjunctive characteristic and if *v* is close to 1 then the operation has a conjunctive characteristic. From this property it can be seen, that by using aggregation we have more possibilities than by using the sum function in FCM. By changing the neutral values at the nodes different operations can be carried out.

5 Components of PCM

The sigmoid function naturally maps the values to the (0,1) interval. Positive (negative) influences can be built with the help of $\sigma_a^{\lambda_1}$, $\sigma_b^{\lambda_2}$ and the conjunctive operator, where $\lambda_1 > 0$, $\lambda_2 < 0$ and a < b.

Using weighted conjunction, it is worth to choose the following values $u = \frac{\lambda_2}{\lambda_1 + \lambda_2}$, $v = \frac{\lambda_1}{\lambda_1 + \lambda_2}$. So we get the generalized positive impulse function:

$$c_m(u,x;v,y) = \frac{1}{1 + \frac{\lambda_2}{\lambda_1+\lambda_2} e^{-\lambda_1(x-a)} + \frac{\lambda_1}{\lambda_1+\lambda_2} e^{-\lambda_2(x-b)}} \tag{11}$$

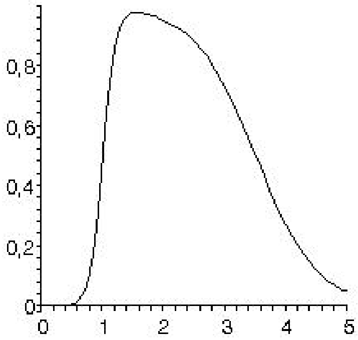


Figure 1
Asymmetrical positive influence on [0, 1]

If the influence is neutral, we represent it by 1/2 value. If there are no influences, then we continuously order 1/2 value to the system. If we want to model positive influences, we order a value, which is larger than 1/2, and maximal value is 1. The negative influence is the negation of the positive influence. To create these influences we use the following transformations:

$$P(t) = \frac{1}{2}(1 + \sigma_{a_1,a_2}^{\lambda_1,\lambda_2}(t)) \tag{12}$$

$$N(t) = \frac{1}{2}(1 - \sigma_{a_1,a_2}^{\lambda_1,\lambda_2}(t)) \tag{13}$$

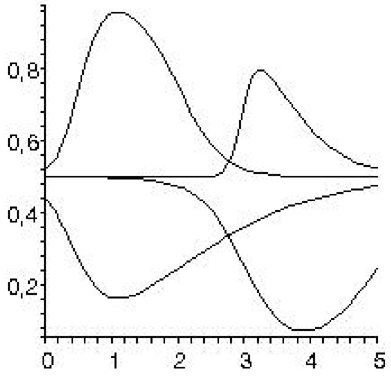


Figure 2
Transformation of 2 Positive and 2 Negative influences

To simulate the system the only thing we have to do is to aggregate the influences. The aggregation operator is a guarantee, that we use influences in the right manner. If we know the real process, then our task is to divide the function values into positive and negative influences. It is an optimalization problem. Starting the optimalization we need initial values. This is the critical point the success of optimalization. If we set it wrong, then the optimalization method finds only the local optimum, but the shape of the real process can help us. We can find the starting and the end points of the influences, even we can estimate the λ parameters. These initial values guarantee that the optimalization method with a high probability finds global optimum. This gives us a new method to analyze real process.

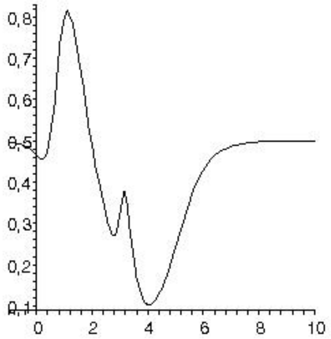


Figure 3
Aggregation of the influences

6 Construction of Dynamic Sytem

With the above mentioned procedure, it is easy to build PCM. The following steps should be carried out:

1. Collect the concepts.
2. Define the expectation values of the nodes (i.e. threshold values of the aggregations).
3. Build a cognitive map (i.e. draw a directed graph between the concepts).
4. Define the influences (i.e. are they positive or negative).

The iterative method:

i Use the proper function or give a timetable for the input nodes.

ii Calculate the positive and negative influences using step 4.

iii Aggregate the positive and negative influences, where the υ_* value is the previous value of C_j.

The system is now ready to make a simulation test. We developed a program in JAVA to test the system. First we study artificial situations. These situation show that the system is very flexible and is easy to adapt to various situations.

Simulation is based on directed graphs. The nodes are illustrated with squares. Between the nodes there are edges. Instead of using arrows, we represent the direction of the edge by a filled circle. If the edge leads from the vertex v to vertex u, then we place the filled circle closer to u. (see Fig. 4). In Fig. 4 an example is given with two nodes and the direction between the nodes is from 2 to 1.

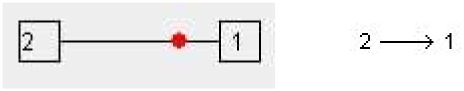


Figure 4
Graph representation

In Fig. 4 index 2 is input node and 1 is inner node. If node is an input node, then instead of giving the initial value we give the input data. There are three types to add new input data: using

- table, we can set input data by our self in every time period.
- Algebraic functions, (*sin*, *cos*, *exp*, *sigmoid*, etc.), calculate the selected function values.

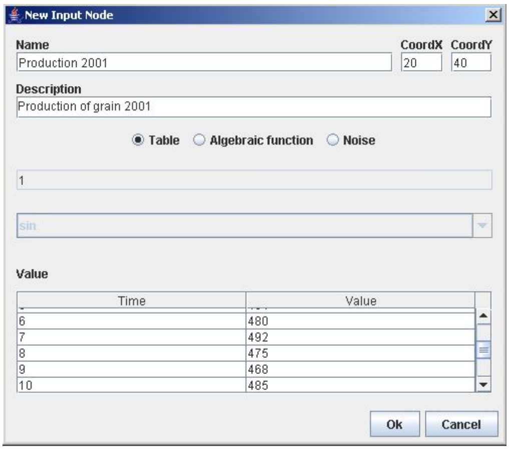
- Generate Noise: we generate random numbers by normal distribution between $[0.5-\varepsilon, 0.5+\varepsilon]$ (default: $\varepsilon=0.1$), which is simulated noise.

See Fig. 5.

The input values are transformed into [0,1] with the sigmoid function:

$$f_{sig}=\frac{1}{1+e^{-\lambda*(x-x_0)}} \ ,$$

where x_0 is the basic value (the expectation level), and λ is the sharpness of the function. It is reasonable to set $\lambda=\frac{4}{x_{\max}-x_{\min}}$, because λ is the slope of the sigmoid function., where $x_{\max}$, $(x_{\min})$ is the largest (smallest) value.



Name: Production 2001

CoordX: 20

CoordY: 40

Description: Production of grain 2001

● Table ○ Algebraic function ○ Noise

1

sin

Value

Time	Value
6	480
7	492
8	475
9	468
10	485

Figure 5
Input node dialog box

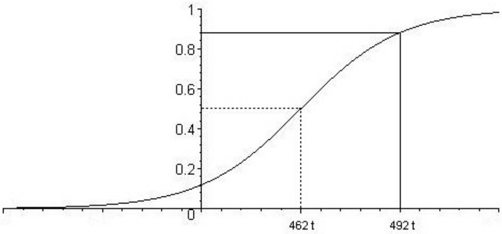


Figure 6
Sigmoid transformation of the food production

The next step is to connect the nodes. To add a new edge, we give

- the index of the source node
- the index of the destination node
- the influence (positive or negative)
- the expectation value(ν).

7 Simulation Process

So we have the system. Now we can start the simulation. In one cycle the following calculation are made:

1 For all edges we calculate the influences.

i We find the source of the edge

ii We transform the source value by the intensity:

$$f_{edge} = \frac{1}{1+\frac{\nu}{1-\nu}\frac{1-node(value)}{node(value)}},$$

where ν is the edge expectation value.

iii We calculate the edge influence with a conjunctive function. If $\lambda = 1$ then the influence is positive If $\lambda = -1$ then the influence is negative. We use the conjunctive function, since in the real world the influences never reach extreme values, i.e.: 0 and 1.

2 Calculate the new value of the nodes.

i We collect the influences that lead to the node

ii We transform the influences, and multiply all of them

$$f_{inftr} = \prod \frac{1-f_{inf(value)_i}}{f_{inf(value)_i}}$$

iii We use the following function to get the actual value:

$$f_{nodenew(value)} = \frac{1}{1+f_{\inf tr}\frac{1-node(value)}{node(value)}},$$

which is an aggregation.

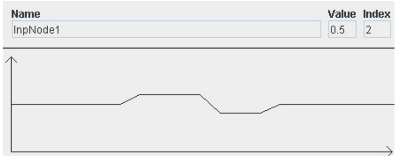
3 Set the actual value of input node.

In the future, for the real world applications we invent learning processes to find the best parameter. This leads to a nonlinear problem.

8 Results

Now we can run the dynamic system and check whether the developed PCM concept fulfills the basic properties.

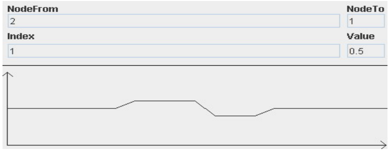
1 We have two input nodes with the same υ value. They always take the same values, but the influences are opposite, one is positive and the other is negative. These two input nodes have one common inner node. The result should be the neutral value (0.5). See Figs. 7, 8, 9, 10, 11.



Name	Value	Index
InpNode1	0.5	2

Figure 7

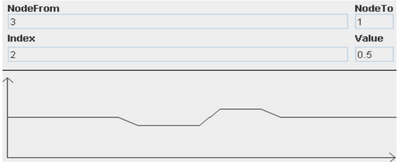
Input node values



NodeFrom	NodeTo
2	1
Index	**Value**
1	0.5

Figure 8

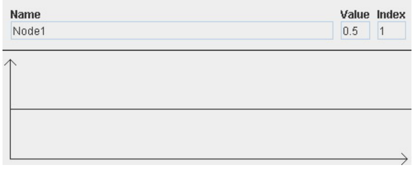
Positive influence



NodeFrom	NodeTo
3	1
Index	**Value**
2	0.5

Figure 9

Negative influence



Name	Value	Index
Node1	0.5	1

Figure 10

Node value, the neutral value

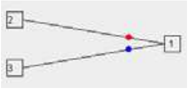
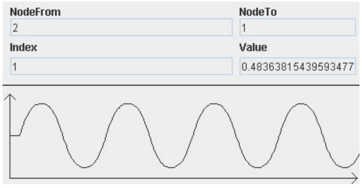


Figure 11

Experiment 1 graph

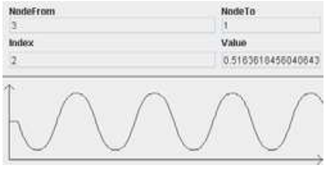
2 Now we have two positive influences with the same neutral values, but the input values are just the opposite (input2(value) = 1-input1(value)) of each other. The result should be also a neutral value. We can also check this by using sin(x) and $\cos(x + \pi / 2)$. See Figs. 12, 13, 14.



NodeFrom: 2
NodeTo: 1
Index: 1
Value: 0.48363815439593477

Figure 12

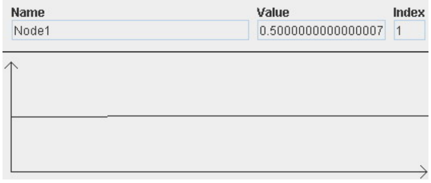
Positive influences of sin(x)



NodeFrom: 3
NodeTo: 1
Index: 2
Value: 0.5163618456040643

Figure 13

Positive influences of $\cos(x + \pi / 2)$

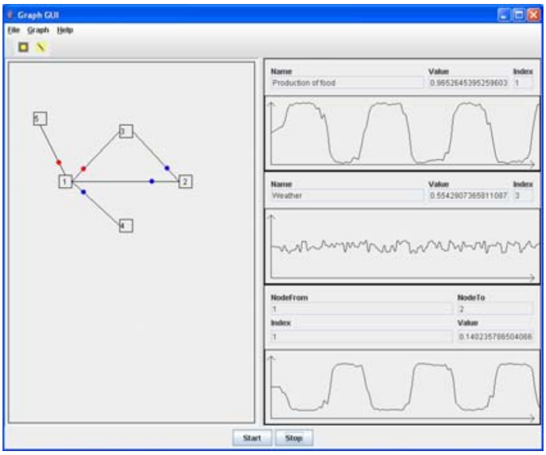


Name: Node1
Value: 0.5000000000000007
Index: 1

Figure 14

The result of the aggregation of two positive influences

3 It is a small complex system with three inputs and two inner nodes. Two input nodes generate noises with different intensity; the third one is a periodical function. See Fig. 15.



Name: Production of food
Value: 0.9652645395259603
Index: 1

Name: Weather
Value: 0.5542907365811087
Index: 3

NodeFrom: 1
NodeTo: 2
Index: 1
Value: 0.140235786504066

Figure 15

Small complex system

Conclusions

We propose a new type of numerical calculus to model complex systems based on positive and negative influences. This concept is similar to FCM, but the functions and the aggregation procedures are quite different. It is based on acontinuous valued logic and all the parameters have semantic meaning. We are working on a real world application and on an effective learning of the parameters of the system.

References

[1] Kosko B.: A dynamic system approach to machine intelligence, Neural networks and fuzzy Systems, 1992

[2] Kosko B., Dickerson J.: Fuzzy virtual worlds, AI Expert, 1994

[3] Kosko B.: Fuzzy Cognitive Maps, nt. Journal of Man Machine Studies, Vol. 24, 1986, pp. 65-75

[4] Dombi J.: Basic concept for a theory of evaluation: the aggregation operator, Europian Journal of Operations Research, Vol. 10, 1982, pp. 282-293

[5] Dombi J.: A general class of fuzzy operators, the De Morgan class of fuzzy operators and fuzziness measures induced by fuzzy operators, Fuzzy Sets and Systems, Vol. 8, 1982, pp. 149-163

Some Aspects of Ambient Intelligence

George L. Kovács [1, 2], Sándor Kopácsi [1, 3]

[1] Computer and Automation Research Institute, Hungarian Academy of Sciences, Budapest, Hungary
[2] University of Pécs and Budapest University of Technology and Economics
[3] Dennis Gábor College, Hungary
gyorgy.kovacs@sztaki.hu

Abstract: Nowadays the competition among companies, joined to the environmental protection rules, is so compelling that they should not only be on the top of technology in they area, but also run their business according to life-long models. The emphasis on the product post-sale life is common for these models. The most popular model is Product Lifecycle Management, for manufacturing companies, or Service Engineering, for service-oriented companies, and, for both, common paradigms are in maintenance, with conformance-to-use certification. The paper introduces basic research results achieved in application of Ambient Intelligence, and suggests considering maintenance as a cross section of the two business paradigms.

Keywords: Product Lifecycle Management, Service Engineering, Knowledge Based Systems, Condition Monitoring Maintenance, Ambient Intelligence.

1 Introduction

In today's highly competitive global economy, the companies should produce high quality products and/or services at lower costs in shorter time (see [1]). This demand forced a number of manufacturing/service industries to apply new strategies for product design, manufacturing and management. For the last decade, the information technology sector made big advance. It made it possible to implement new supply chains, delivering *products-services*, supported by co-operating organizations. This type of changes opened horizons to enable new business paradigms, such as product lifecycle management or service engineering, embedding high-intensity information flow, with enhanced transparency of the material resources decay and with increased intangible value added. These paradigms make it possible for networked companies to be more competitive in a novel business area, improving after-sale service, products maintenance and recycling. In some cases, to excel in service and maintenance area is the critical means for a company to be winner on the market. Indeed, the full acknowledgement of the technical conformance of a product at the *point-of-use* is a non eliminable request, to comply with eco-consistency incumbents, and the reliable recording of the lifecycle falls-off becomes standard duty to certify the item sustainability achievements.

This paper will give some details on most necessary disciplines and programming tools and methods to understand and realize a very characteristic example: predictive maintenance.

2 Some Corresponding Production/Business Paradigms

Recently, manufacturing world was moving from an economy of scale to an economy of scope, under the global economy for customers' satisfaction (see [1], [3]). Under that conditions, for most companies around the world, surviving in business means to satisfy three challenges, they are: to meet customer requirements; to reduce the time-to-market of their products; and manufacture products at lower environmental impact. These changes in industry have been reflected on human society. Following [3], the earlier affluent society, needs be turned into a more conservative thrifty society, as the world, we are living in, cannot deal with ceaselessly consuming raw materials.

Of course, as well known (see [1]), goal in the industrial economies is value creation by marketing products. Nowadays, in many businesses, the created value consists of many components and the weight of the value related with tangible product turns to be but a fraction of the whole delivery. In the thrifty society, frugality and recycling play a relevant role; the traded artifacts are replaced by products-services, supported by extended enterprises (see [3]). In this type of manufacturing organizations, the value is added by the supplier on the lifelong steps of the provision, aiming at new business paradigms types. This type of business organization has been acknowledged, and the, so called, Product Life Cycle Management, PLM, tools are example achievements.

This paradigm makes possible the manufacturing set-ups complying with the thrifty society demands, due to the transparent recognition and control on resources decay and environment burden. The previous challenges for companies add or extend over new ones. If formerly meeting customer satisfaction meant to develop product with high performance and quality, nowadays companies need, moreover, supply lifelong services and maintenance procedures, including dismissal and recycling incumbents.

All the above-discussed issues led to new rules on the market:

Rule 1. It is not enough to produce the required product, but the most important is to produce after-sales services, which satisfy or, in some cases, predict, as well, the customer requirements, complying with the enacted environment protection rules.

Rule 2. Reducing the time-to-market period of the product means the reduction of the time-to-market period of the services, supplied as integral part of the delivery, with due concern on the technological sustainability of the traded goods.

Rule 3. Not only goods should be developed to satisfy demands of low cost and eco-consistent quality, but accompanying services shall, as well, satisfy these demands, further providing full visibility of the impacts on the human natural surroundings.

The farther our community will turn from the affluent to the thrifty society, the more relevance will be given to the products related services in companies' competition. In other words, if in the affluent society, competition between companies producing *product-service* deliveries made emphasis on "product", in the future, in the thrifty society, the competition will make emphasis on the "service" part of the delivery. But it does not mean that the role of the product component could be totally replaced.

Still, the role of the second part of the product-service supply is getting more relevant and companies could build-up their businesses in the area of providing accompanying services for products manufactured by another company. The context, thereafter, has recently, led to the appearance of a new scientific field, say, Service Engineering, SE, purposely developed to expand the (mainly) intangible provisions for enhanced use of conventional goods.

2.1 Product Life-Cycle Management (PLM)

The Product Lifecycle Management, PLM, concept appeared in the second part of the 1990's. This concept provides a platform to share product related knowledge across an extended enterprise, from product design and creation, through dissemination and after sales services, up to product dismissal and recycling. Following [2], PLM is defined as *"a new integrated business model that, using ICT technologies, implements an integrated cooperative and collaborative management of product related data, along the entire product lifecycle, dismissal included"*.

The right hand side of Figure 1 shows a graphical interpretation of the PLM concept. As one can see, the PLM joins three main chains of the extended enterprise: Engineering chain, Operation chain and Support activities chain. Any chain consists of sub-processes. For example, Engineering chain contains three sub-processes: product design, process planning and factory planning.

It is possible to highlight several reasons which led to build a PLM. From the manufacturer viewpoint, product innovative up-grading, customer-driven quality, operation excellence, etc., require enhanced visibility on actual lifecycle data. From the outer market viewpoints, items responsibility at the point of service, product complexity, shrinkage in the duty life, dismissal incumbents, etc., push toward higher transparency of supply chain and environmental issues.

The implementation of the PLM concept is impossible without proper ICT tools. F. Ameri and D. Dutta in they work [4] summaries the impacts which PLM gives to ICT solutions applied in manufacturing. Let us enumerate some of them:

- It makes a closer connection between, on one side, engineering and manufacturing and, the other side, finance and marketing, assessing the criticality of the design steps for the after-sale services.
- It provides a unified information model to take into account all relevant facts throughout the delivery lifecycle, assuring data-vaulting organisation and recording, fulfilling all the needs of the direct or indirect stakeholders.
- It fills the gap between enterprise business processes and product development processes. Following [4], PLM works as a glue which adhere all the processes that have something to do with product and connects all functional silos to make them horizontally integrated.

The build-up of PLM solutions requires to re-engineer the company's business processes and to implement the supporting ICT solutions. As concluded in literature (see e.g. [2]), PLM links together, not only processes at the engineering, operation management levels with non-production services (like marketing, sales, after sales, quality or maintenance), but, also, the supporting ICT solutions. The management of maintenance activities is often integrated into Product Lifecycle Management (PLM) tools (see [8]). The two main reasons for using PLM tools are process integration and data integration. Therefore, they usually consist of a workflow and a data integration component (see [8]).

2.2 Service Engineering (SE)

Although, methodologies related with organizations providing services are covered in PLM, it is more focused on integration of services with manufacturing processes. According to that so-called Product-Service supply can be developed, where the emphasis can be either on Product part or on the Service part (see Figure 1). If the importance is on the Product part, the model is fully applicable for several business cases but, from the authors' point of view, for the non-manufacturing business the discipline, called Service Engineering (SE), is more appropriate. This business paradigm makes the stress on Service part of Product-Service supply (see Figure 1).

Service Engineering is a new research discipline (see [5]), which mainly deals with improving design process of the service, developing, implementing services design, development and execution as corporate function. At last, but not least, the service engineering has to deal with service-oriented human resource management.

In some aspects Service Engineering is very similar to the PLM concept. Service Engineering assumes that services can be designed and re-developed in a similar manner as physical products. Based on [5] the process of service design and development has three major phases, which include some set of sub-processes. These are: service planning, service conception and service implementation. The graphical interpretation for SE is given on right-hand side of Figure 1.

One example of the ICT solutions for the tasks considered in SE are e-maintenance systems. These systems can vary from product supported or from engineering/business field of application. Like PLM solutions their may consist of different modules and tools. Some modules can evaluate products, machines or other equipment without human intervention or adjust monitored objects to avoid breakdowns or undesirable situations. If remote support is not effective such systems can generate instructions or automatically call service team on site. Such maintenance systems can use knowledge-based solutions or modules and tools to provide multimedia based features for collaborative maintenance.

2.3 Extended Products

As the term extended enterprise comprises more than just a single enterprise the term extended product [23] should comprise more than just the core or tangible product. The view of the EXPIDE project of the EU [24] works on extending a formerly tangible product. For that reason different services in an intangible shell around the tangible product should be the focus. The explanation for this idea is depicted in a layered model (concentric rings) that can be applied to various types of products.

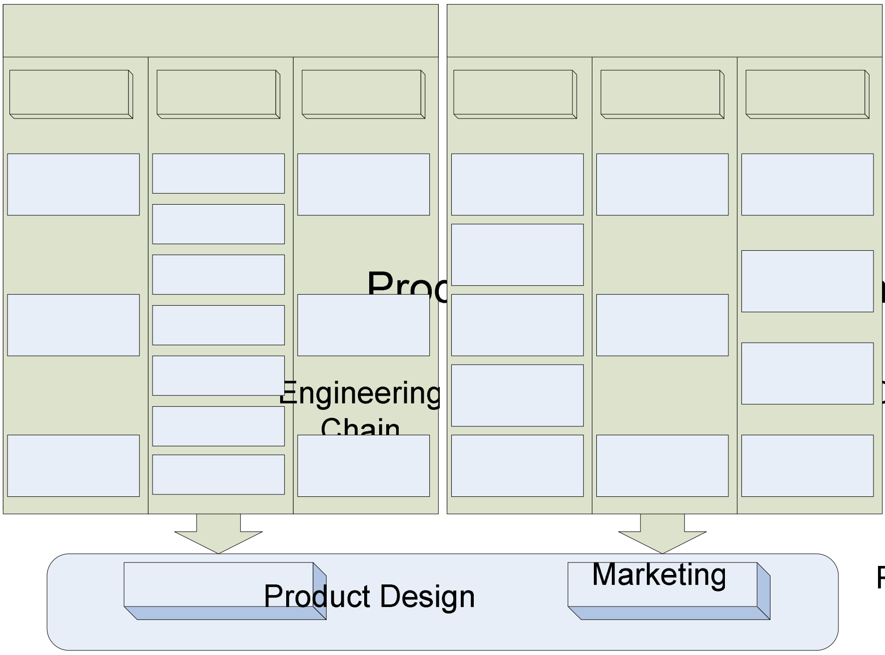


Figure 1
PLM and SE business paradigms

The layered model was developed in order to structure the extended product approach. This model shows a type of hierarchical, physical extension of (extended) product services. The concept with three rings can be described as follows:

- There is a core product that is closely related to the core functions of a product. As an example, this approach can be taken to characterize the various types of shapes of a part: Shapes that are relevant for the functionality and others that are not that relevant i.e. shapes where the freedom of the designer to define tolerances, etc. is higher. Consequently core functions of a car are parts like engine or wheels.
- The second ring describes the packaging of the core functions. The second ring only includes tangible features of the chosen product. The features of the tangible product are different from manufacturer to manufacturer (supplier to supplier). In the car manufacturing industry we have Mercedes/Porsche and Daihatsu/Fiat which both supply cars but are they doing this in a similar fashion? The answer is clearly "no".

In spite of the above given logical explanation the distinction between core product and tangible product is not trivial and not (yet) commonly accepted.

- The third ring summarizes all the intangible assets of the product. Intangible assets surround the tangible product. In general they could be the same for similar products. However, in practise they can be quite different, for example, if we compare the Mercedes/Porsche and Daihatsu/Fiat service strategies. While Mercedes/Porsche are very concerned with the customers' demands, Daihatsu/Fiat are continuously concentrating on reducing the costs in order to offer the cheapest possible cars to the customer.

We may say that the extended product of EXPIDE [24] includes the following elements:

- A combination of a physical product and associated services/enhancements that improve marketability.
- Tangible extended products, which can be intelligent, highly customized, user-friendly and include embedded features like maintenance.
- Intangible extended products, which are information and knowledge intensive and can consist of services, engineering, software, etc.

Customers are focusing on the benefit out of a value-adding service and not anymore the physical product itself. One can make a difference between products in a narrow sense and a product in a broader sense. By narrow we consider the product as a tangible entity which is offered in the market whereas the broader sense gives an indication about the objective of the product which means solving a problem of the customer or satisfying a demand. Services of the third ring may be services provided by the product and/or services necessary for the product. This concept of EXPIDE concerns mainly the services provided by the product.

In the following, where we take into account product lifecycle and service engineering we are concerned about services which are necessary for the lifecycle of the product, unlike the EXPIDE project which extends the product mainly by services provided by the product.

3 Maintenance - a Link between PLM and SE

Maintenance activity usually included in PLM paradigm and in Service Engineering as well. In other words maintenance is an overlapping area for these two paradigms (see Figure 2 for a simplified view).

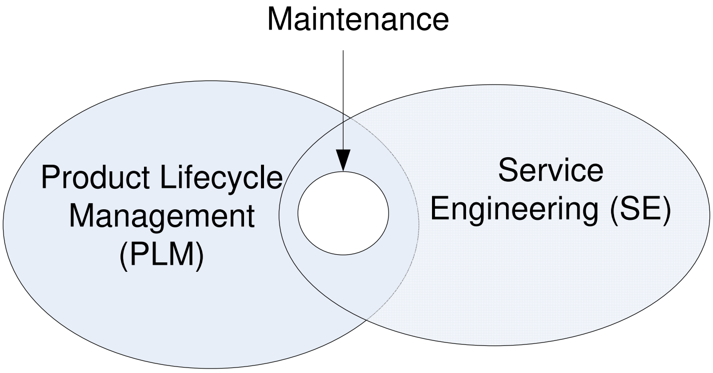


Figure 2
Maintenance is cross-paradigm area

There are several types of maintenance strategies, commonly in use today. The first one is breakdown maintenance (in some cases called as corrective maintenance, see for example [6]). Alternatively more elaborated strategies can be devised, which partly give possibility to avoid the failures by preventive (or planned) maintenance, either, to detect the symptoms of anomalous running and to enable reactive or proactive maintenance operations, before actual failures develop. Just a list of maintenance strategies commonly used:

- Breakdown/Corrective Maintenance
- Preventive Maintenance

On the base of [6], three different approaches of the preventive maintenance are mentioned. There are systematic/scheduled, conditional and predictive maintenances.

  - Systematic/Scheduled Maintenance
  - Conditional Maintenance
  - Predictive Maintenance - e-Maintenance

3.1 Knowledge Based Systems

Knowledge-Based Systems are computer programmes that use knowledge about some domain to reach a solution for a problem of that domain. This solution is essentially the same as that would be concluded by a person knowledgeable enough about the domain of the problem when he were confronted with the same problem. Knowledge based systems allow us to separate the knowledge and the search/inference engine.

Knowledge based systems are capable of analysing massive amounts of data on a subject, applying reasoning ability to that data and providing answers to a given problem. Their purpose is not so much to remove the need for having experts in a particular field, but to provide the ability of an expert in circumstances where a human expert is not available or not efficient. In condition monitoring maintenance, for instance, the inclusion of heuristic blocks is very effective to work out diagnostic frames based on complex signatures and very large reference features.

3.1.1 Reasoning Methods

Reasoning methods and tools are providing ways of Knowledge Management from knowledge capturing to re-use for reasoning. Generally, the complete lifecycle of knowledge management is supported by different ways, say:

- To capture knowledge in different forms.
- To store the knowledge.
- To provide mechanisms to re-use this knowledge for reasoning on 'similar' problems.
- To maintain the knowledge.

There are several reasoning methods covered by research activities. Analysing and comparing the reasoning methods, the following statements can be made:

- Deductive and probabilistic reasoning are, respectively, an enhancement and a specialisation of rule-based reasoning. These methods require the definition of rules, which require considerable effort and a good knowledge of the domain.
- Causal and temporal reasoning require other reasoning methods: as they represent a link and enhancements to other methods, they are not stand-alone.
- Frame-based reasoning requires an extensive knowledge of the domain and has a difficult maintenance.
- Case-based reasoning uses only the knowledge collected during the system normal functioning, structured into cases.

3.2 Agent Based E-maintenance Systems

The agent based e-maintenance systems are alternative to the centralised KBS approach. In this case, an intelligence concentrated in KBS is distributed between agents, specialised to monitor the different parameters. The benefit of such systems is that the agents can be technologically oriented on the monitored product, communicate to each other and provide not data, but higher-order information, to the KBS system.

The example of agent-based solutions can be found in works of Lee [9]. Lee applies such solutions for execution predictive maintenance of industrial machine tools. But it can be applied for other product-service delivery, too.

The core-enabling element of an intelligent maintenance system is the smart computational agent that can predict the degradation or performance loss, not the traditional diagnostics of failure or faults. In case of [9], these agents are called as "Watchdog Agents" and can be described as mechatronical devices with embedded computer (with software agent), which provide the intelligence. Agents may transform data to information, and information to knowledge and synchronise decisions with remote systems, contain embedded prognostics algorithms for performance degradation assessment and prediction. A product's performance degradation behaviour is often associated with multi-symptom-domain information cluster, which consists of degradation behaviour of functional components in a chain of actions. KBS can be efficiently applied to handle multi-symptoms.

But the main drawback of the system described in [9] and many other agent-based solutions available on the market is that their high level system automation possibilities and human oriented interfaces are not incorporated enough. The situation can be changed drastically by implementing Ambient Intelligence (AmI) concepts.

4 Ambient Intelligence – Enabling Technology for Maintenance

4.1 The Concept of Ambient Intelligence (AmI)

The concept of Ambient Intelligence (AmI), relies on provisioning of *ubiquitous computing* (i.e., useful, pleasant and unobtrusive presence of computing devices everywhere), *ubiquitous communication* (i.e., access to network and computing facilities everywhere), and *intelligent user adaptive interfaces* (i.e., perception of the system as intelligent by people who naturally interact with the system that automatically adapts to their preferences).

Ambient intelligence is a rapidly increasing field of information systems that has potential for great impact in the future. The term "ambient" is defined by Merriam-Webster's dictionary [14] as "existing or present on all sides". The term Ambient Intelligence was defined by the Advisory Group to the European Community's Information Society Technology Program [15] as "the convergence of ubiquitous computing, ubiquitous communication, and interfaces adapting to the user" [16]. Ubiquitous should also be defined since the core domain of AmI envelops this concept. Ubiquity involves the idea that something exists or is everywhere at the same time on a constant level, for example, hundreds of sensors placed throughout a household, or in a factory where some number of agents combined into the network which can monitor the operation of household equipments, machine tools or the production of any future product. This idea is important when trying to understand the future implications that AmI will have on the environments we live and function in.

Ambient Intelligence incorporates properties of distributed interactivity (e.g. multiple interactive devices, remote interaction capabilities), ubiquitous computing (the “invisible” computer concept), and nomadic or mobile computing. Ambient Intelligence has the potential to provide the user with a virtual space enabling flexible and natural communication with the computing environment or with other users, providing input and perceiving feedback by utilizing proportionally all the available senses and communication channels, while optimising human and system resources [13].

The objective of AmI is to broaden the interaction between human beings and digital information technology through the use of ubiquitous computing devices. Conventional computing primarily involves user interfaces (UIs) such as keyboard, mouse, and visual display unit; while the large ambient space that encompasses the user is not utilised as it could be. AmI, on the other hand, uses this space in the form of, for example, shape-, movement-, scent- and sound-recognition or -output. These information media became possible through new types of interfaces and will allow drastically simplified and more intuitive use of devices. Wireless networks will be the dominant technology for communication between these devices. The combination of simplified use and their ability to communicate will eventually result in increased efficiency for users and will, therefore, create value, leading to a higher degree of ubiquity of computing devices. Examples of such devices range from common items such as pens, watches, and household appliances to sophisticated computers and production equipment.

4.2 Ubiquitous Computing

Mark Weiser [17] coined the term "ubiquitous computing", referring to omnipresent computers that serve people in their everyday lives at home and at work, functioning invisibly and unobtrusively in the background and freeing people to a large extent from tedious routine tasks. The general working definition of ubiquitous computing technology is any computing technology that permits human interaction away from a single workstation. This includes pen-based technology, hand-held or portable devices, large-scale interactive screens, wireless networking infrastructure, and voice or vision techno [18].

In its ultimate form, ubiquitous computing means any computing device, while moving with you, can build incrementally dynamic models of its various environment and configure its services accordingly. The devices will be able to either "remember" past environments they operated in, or proactively build up services in new environments (Lyytinen and Yoo 2002).

Ubiquitous computing is roughly the opposite of virtual reality. Where virtual reality puts people inside a computer-generated world, ubiquitous computing forces the computer to live out here in the world with people [17].

4.3 Ubiquitous Communication

Today, numerous objects are equipped with computers, i.e., our environment already exhibits a relatively high level of ubiquitous computing. However, in most cases, the computers do not operate at their full potential since they are unable to communicate with each other. A major change in the corporate and home environments that will promote ubiquitous communication and, thereby, ubiquitous computing is the introduction and expansion of wireless network technology, which enables flexible communication between interlinked devices that can be stationed in various locations or can even be portable. To implement wireless technology on a wide level, however, the wireless hardware itself must meet several criteria on the one hand, while easy integration and administration as well as security of the network must be ensured on the other.

The agent technology is one of the enabling technologies for realization of AmI concept in life. Following [7] it is possible to say that agent technology will provide new distributed architectures and better communication strategies for the applications, making easier the information exchange and allowing to integrate new modules like sensors or diagnosis algorithms with less effort from customer and machine tool builders' point of view.

4.4 User Adaptive Interfaces

User adaptive interfaces, the third integral part of AmI, are also referred to as "Intelligent social user interfaces" (ISUIs) [19]. These interfaces go beyond the traditional keyboard and mouse to improve human interaction with technology, by making it more intuitive, efficient, and safe. They allow the computer to know and sense far more about a person, the situation the person in it, the environment and related objects, etc., than traditional interfaces can.

ISUIs encompass interfaces that create a perceptive computer environment, rather than one that relies solely on active and comprehensive user input. ISUIs can be grouped into five categories:

- Visual-recognition (e.g. face, 3D gesture, and location) and -output
- Sound-recognition (e.g. speech, melody) and -output
- Scent-recognition and -output

- Tactile-recognition and -output
- Other sensor technologies

Traditional user interfaces like PC-controlled touch screens in a company environment and user interfaces in portable units such as PDAs or cellular phones can also become ISUIs. The key to an ISUI is the ease of use, say, the ability to personalise and automatically adapt to any particular user behaviour patterns (profiling) and actual situations (context awareness), by means of intelligent algorithms. In many cases, different ISUIs, such as voice recognition and touch screen, are combined to form multi-modal interfaces [20].

Interfaces, especially user interfaces are one of the crucial building blocks for AmI, because they define the experience the user will have with the intelligence surrounding him/her. Major importance is attached to the so-called natural-feeling human interfaces and to multimodal interfaces. Vision technologies and displays are part of the interfaces as well. Breakthroughs in user interfaces are important for consumer acceptance, not only, for the mass markets in general but also for people with disabilities in particular. The vision of AmI assumes that the physical interaction between humans and the virtual world will be more like the way humans interact in the real world, hence the term natural interfaces. Humans speak, gesture, touch, sense and write in their interactions with other humans and with the physical world. The idea is that these natural actions can and should be used as explicit or implicit input to AmI systems. Interfacing should be completely different from the current desktop paradigm (GUI-Graphical User Interface) based on keyboard, mouse and display. This also means that interfaces should be multi-model, as humans communicate in a multimodal way against machines that typically operate in a single mode. Moreover, with current interface technology humans must learn and understand the computer language; in the future, this would be the other way around.

We should also consider the intelligent interfaces between the components (i.e. products, production units) in a shop floor environment. Machines and products should communicate with each other in order to realize the intelligent behavior of the system. This kind of communication requires other types of adaptive interfaces.

4.5 Some Application Fields and AmI Drivers

The AmI vision anticipates that ICT will increasingly become part of the invisible background to peoples' activities and that social interaction and functionality will move to the foreground resulting in experiences that enhance peoples' lives. From this anticipation, it is clear that AmI could and should be found everywhere in the human environment.

The realization of AmI concepts will lead to establish a collaborative working environment, where virtualized entities will communicate to each other. These entities can be (see [11]) humans, artificial agents, web/grid services, virtualized-entities representing the real things (not only human beings), descriptions of human knowledge (knowledge based systems) etc. Following [11] these entities will be able to interact with one another in an AmI environment to leverage the full potentiality of network-centric environments for creativity improvement, boosting innovation, and productivity gains.

Such networks will provide possibility for individuals to experience interaction with human and artificial agents in their working environments.

AmI by the means of ubiquitous computing and communication technologies give great impact on future maintenance technologies and business paradigms like PLM and SE, as well.

4.5.1 AmI PLM

The main advantage of PLM solution for a company is that it integrates many enterprise IT solutions in one system, and, as result different processes, can be synchronised. Let us determine AmI PLM system as a system that supports and executes an integrated cooperative and collaborative management of product related data, along the entire product lifecycle by realising ubiquitous computing, ubiquitous communication and intelligent user adaptive interfaces. The implementation of AmI concepts to support PLM business paradigm will drastically change PLM systems. Different processes from different chains (see Figure 1) will be integrated in one virtual environment. The component of this environment will be entities which can be humans, IT systems (as ERP), intelligent software or mechatronical agents. These entities will communicate on peer-to-peer way and build up product oriented dynamic networks. These networks are dynamic because on different steps of product life-cycle different entities participate in the product oriented network. The example of such a network is represented on Figure 3. The Figure shows an example of a network on the step of generating idea about a product. Two types of graphical representation for entities are used. One is a two dimensional human sign for human entities and another –cycle with line under it-for software agent. The communications between entities are shown as lines with arrows. It is clear that depending on the product or step in the product life-cycle different entities will join or leave the network.

The implementation of ambient intelligent concepts and integration it into PLM business paradigm will lead to appearance of new paradigm and new environment: the Ambient Intelligent Product Lifecycle Management (AmI PLM). The main benefits of such an environment will be the following:

- Concentrating information around product and creation product oriented environment for all steps of product life-cycle
- Enabling information sharing, easier access and management of the product related data

- Product information updates performed in real-time and in intelligent ways
- Product life-cycle management is getting easier
- Enabling products to carry and process information, which influences their destiny…

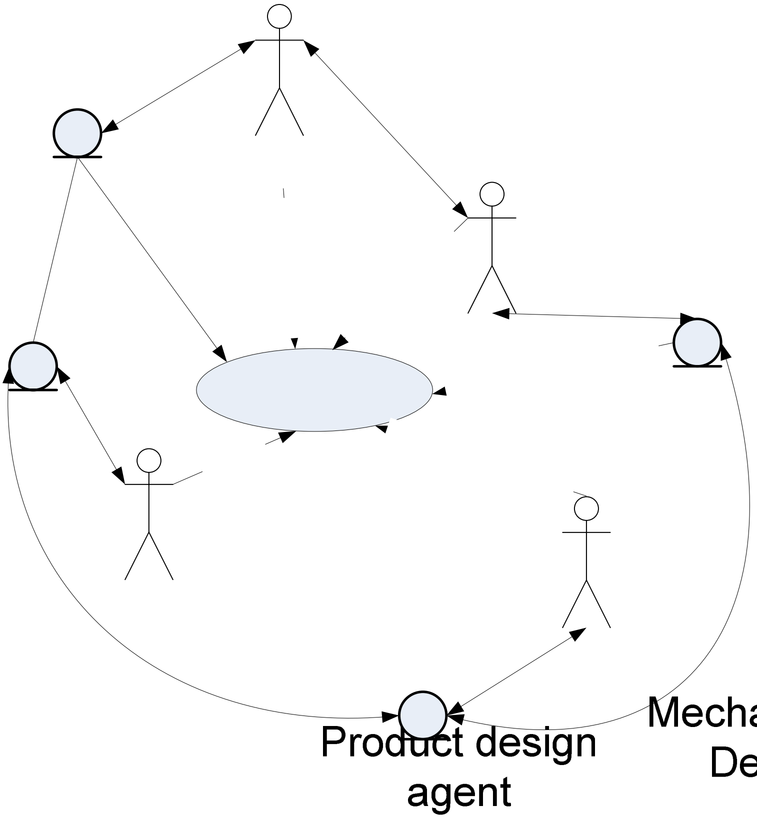


Figure 3

AmI empowered product oriented network

4.5.2 AmI SE and Maintenance

The impact of AmI on service providing and service engineering will be similar to impact on PLM. The realisation of the AmI vision will require the development of large, complex, heterogeneous, distributed systems. These must be built on heterogeneous platforms capable of providing seamless networking so as to support the delivery of layers of value added services or functional services to the individual, to industry, and to administrations. The resulting systems, comprising several interacting embedded software components, will need to be intelligent, self-configuring, self-healing, self-protective, and self-managed. It will lead to appearance of new type of services - Ambient intelligence services.

As it was discussed above, the management of maintenance activities is often integrated into Product Lifecycle Management (PLM) tools, or it can be some kind of service supported by independent ICT solution.

Maintenance tasks tend to be difficult (see [7]), they require expert technicians. Maintenance working conditions are characterised by information overload (manuals, forms, real-time data ..), collaboration with suppliers and operators, integration of different sources of data (draws, components, models, historical data, reparation activities).

Ambient Intelligence will provide a new working environment to maintenance technicians offering: access to ubiquitous and up-to-date information about the equipment wherever the equipment or the operator is (enabling remote maintenance and life-cycle management) and user friendly and intelligent interfaces (context-aware applications).

Advantages provided by the use of AmI in maintenance environment come from:

- Simplifying distributed computing: better distributed knowledge management.
- Intelligent resource management.
- Overcoming user interface problems.
- Overcoming data exchange and communication problems.
- Personalization, adaptation to the user.

The implementation of AmI principles in maintenance will lead to appearance of new type of maintenance - Ambient intelligent maintenance - and will give large possibility for realisation self-maintenance.

5 A Demonstration Project

One of our test beds for our ideas was the European Research Project, called FOKSai (COOP-CT-2003-508637, [21]), of which the acronym stands for “SME Focussed KM System to support extended product in Ambient Intelligence domain”.

The main goal of the project is the development of a knowledge management (KM) system as an extension to Ambient Intelligent products for SMEs in four industrial areas. This main goal, common for the four consortium SMEs, will be achieved through the development of a sophisticated support system to the extended AmI-products of the companies. All these SMEs plan to introduce in the near future (next 1-2 years) new and/or improve their current products with even more AmI features seeing this as their crucial competitive edge. A considerable enhancement of business performances of the four SMEs will be reached by introduction of the FOKSai system mainly by reduction of time and costs for customer support (such as product maintenance, solving customer problems, etc.). The new support system should result in a significant shortening of down times of the AmI products and cutting of the maintenance costs.

5.1 Expected Results

Technically seen the project will develop:

- a methodology for extension of AmI products which will strongly observe business and organizational issues relevant for SMEs,
- a knowledge-based system to support extended AmI-products, which will be affordable for the SMEs.

The FOKSai solution is planned to be general enough to be applicable for different products and scalable to support future AmI products in order to achieve a product (methodology and knowledge management system), which can be offered to a wide spectrum of SMEs intending to introduce AmI products in their product portfolios.

The topics of customer and product support to be developed and demonstrated as pilot installations within the environments of four SMEs in the project consortium include:

- remote supervising, problem identification, problem solving,maintenance of heterogeneous customer systems, and subsequently reduction of efforts/costs for searching of the reasons of problems in products containing AmI components,
- e-supporting manufacturer's staff at remote customer site location to solve customer/product problems,
- proper integration and sophisticated knowledge-based interpretation of the intelligent ambience information and "reactions",
- gathering and structuring of the AmI-product and process knowledge, from the problem solutions, for the reuse in innovations introduction,
- direct feedback from user to AmI-product/service design and development.

5.2 Main Concepts of the FOKSai Project

A generic concept of the system was identified based on the requirements of the end-users. According to this basic concept the system comprises the following modules (See Figure 4):

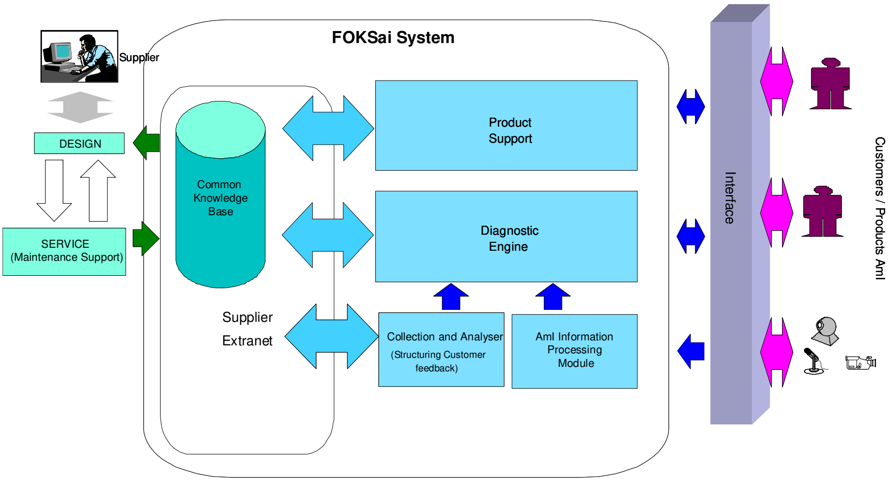


Figure 4
FOKSai System Concept

- *Common Knowledge Base* (CKB): The central repository of the product related knowledge, containing the data and information on problem solving, knowledge on AmI features in products, results of analysis of AmI information and product reference information.
- *Product Support*: This module includes information, documents and knowledge relevant for product and customer support (e.g. information on new product models, new services, advises how to apply products and services in order to avoid problems etc.).
- *Diagnostic Engine*: The central KM based module providing interactive problem solving assistance to users. This module, based on the proven case-based reasoning method, allows to quickly get the required problem solving suggestion. The interactive features also allow user to add useful information – the system 'learns' as the knowledge base grows.
- *Collection and Analyser*: This is an interactive tool to build the knowledge base, and enabling the user to organise the information in it to suit optimally the decision making support by problem solving. This module provides the facilities for the users to analyse the knowledge in the knowledge base, as well. The users can analyse the most frequent problems and to decipher customer feedback regarding products or services, specifically regarding AmI features.
- *AmI Information Processing Module*: This module processes information from the AmI parts of the product as an input for the diagnostics support and for the Common Knowledge Base.

5.3 Business Cases

Based on the system concept the business cases in the project partner companies have been identified and the user requirements have been specified, that have resulted in detailing of the system modules and relations among them and to the legacy systems. These industrial applications are from four different SMEs with different profiles from four European countries (Germany, Great Britain, Spain and Hungary). The dissimilarity ensures, that the results of the project can be later widely used.

The objective of the first Business Case (RegioData, Germany) is to establish a new form of customer support by implementation of a new methodology and ICT system to support the already introduced extension in the form of remote maintenance system, which is the step further in the customer support. The new system should include not only system status and ambient related data but also the data illustrating the state of the (running) system performance, obviously necessary for the achievement of the full system availability and advanced maintainability. The knowledge-based remote access maintenance system will provide a very fast reaction in the case of customer problems. Especially in a case of real-time critical business or process disturbances requiring immediate corrective actions, the efficient action via such a remote access system is of a major importance to achieve customer satisfaction and to increase his trust. Intention is to install Internet-based knowledge system for remote preventive and upon request maintenance and provision of general after sales services including the support of ambient intelligence systems. The knowledge system, which will be the base of the new system, will include all the problems previously detected and the solutions identified, helping to solve the new problems and quickly react in the case of problems re-occurrence. In addition, the information from ambient intelligence system will be introduced. The currently used facilities and data (e.g. customer support from the call centre) will be reused. For example, they will include appropriate features to assess the process critical parameters and indications of the process critical states, based on the collected and properly structured knowledge.

In the second Business Case (LANeX, Hungary) the need of very high data security providing the users trusts in such a system is obvious. Taking into account the number of measurements and information, from the intelligent ambient and network, which are analysed continuously as well as their mutual dependencies, it is clear that a rather complex KM system is necessary for the proper remote product maintenance. The application of the knowledge-based and web-based FOKSai system will have to provide the quick reactions to any disturbance, taking into account the importance of the information processed. The extensions to the products, connected to its monitoring and control system, have to be scalable in order to offer a high variety of services. A pro-active customer support will be needed, offering to the operator an online help and consultancy function. The pro-active concept should enable also a collection of feedback from customers in order to improve services continuously. The feedback from customers will be analysed and forwarded to design department as well, aiming at continuous innovation of services and equipment. One of the main requirements of LANeX upon FOKSai system is the *scalability* and *versatility* of the functions it should fulfill, taking into account the variety of the users (of apartments and offices) and their needs.

The third Business Case (DISTEC, Spain) focuses on several activities, which can be mainly summarized as control engineering, technical process automation, tele-management, studies and technical projects. Tele-management can be defined as the remote management for technical installations, which allows the monitoring, management and actuation (when it is required). Furthermore, it can also be applied to different areas, such as chemistry, metallurgy, textile industry, potable and residual waters, control of refrigerator chambers, electrical mini-centrals, etc. Telemanagement provides security and comfort to every user, from the installation responsible to the manager, without forgetting the maintenance technician. This technique assures a permanent monitoring of the installation and immediately alert the responsible person by means of the remote-alarm warning. The intervention is fast and the displacement only is required in extreme conditions. All these advantages widely reduce the effort in case of system exploitation and maintenance.

The SME in the fourth Business Case (CTOOLS, UK) produces process automation and cutting machinery and supplies other companies with turn-key solutions. Installation and set-up support as well as problem solving support provided on-line (mainly per telephone), either to own field engineers or to customer maintenance staff, is one of the most important activities of CTOOLS. CTOOLS is though often requested to send engineers to solve problems at the customer's site as well. This 24 hours support, in either form, involves rather high costs. The company needs to support their products in more efficient and cost-effective way, i.e. it needs a way of providing problem solving help to customers (and field engineers) in order to reduce the resources required to support their products and to increase the efficiency of the maintenance teams. Increasingly complex CTOOLS machines (products) require a number of built-in measurements and control of the correct performance of the machines. The use of knowledge based analysis of ambient data will enable adjustments to be made when they are required. For example, if the temperature is too low (as decided by the knowledge based analysis), then the system will inform the customers that they need to increase the environmental temperature by a number of degrees, and not to do any further cutting until this temperature is reached, otherwise the cutting performance will be poor. The appropriate form of knowledge from the FOKSai system will be provided to the machine operators who will take the necessary corrective action to adapt the environment to the required conditions.

5.4 An Early Prototype

As described in section 0, the FOKSai system is divided into several modules (see Figure 5). The FOKSai early prototype, that has been elaborated by the current phase of the project, comprises these modules with restricted number of the functionalities realized, as it can be seen from the modules description below.

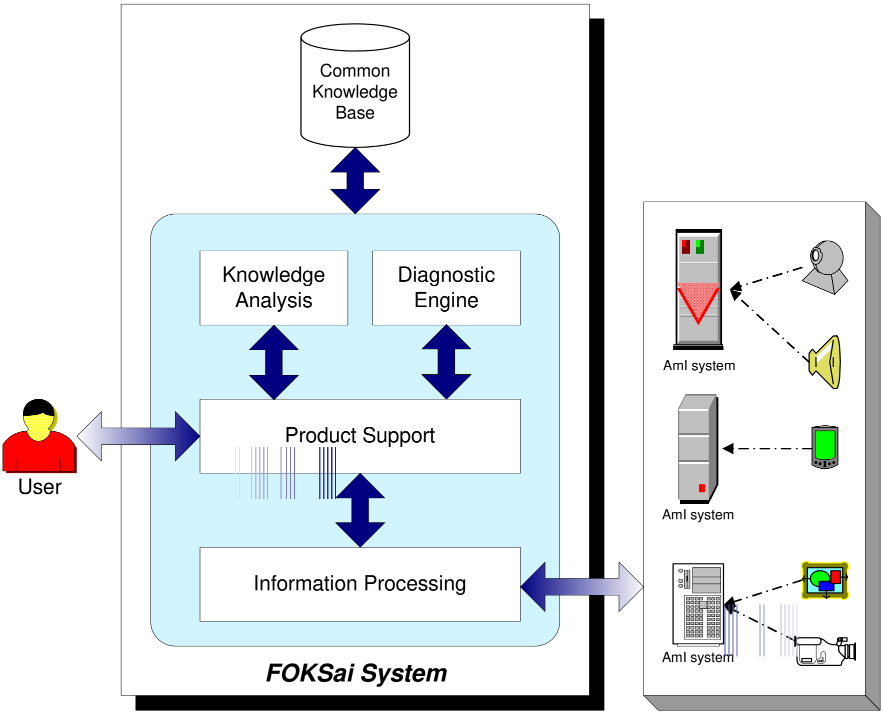


Figure 5
FOKSai modules

Common Knowledge Base

The Common Knowledge base (CKB) has the objective of storing all the information that describes products and processes, as well as all the necessary information related to these components. This repository was implemented at the end users as a relational database, using Oracle or MySQL, depending on their individual requests. Further completion of the data in CKB will be done during the full system prototype development, but no refinements and changes are planned in the CKB structure.

Set-up Module

The Set-up Module is a clear and efficient graphical user interface, that enables the users to understand the meta model and make the best use of it. This Module has been realised in the scope of the early prototype as a stand-alone java application, installed at the companies. This module supports the definition, modification and deletion of all the information that constitutes the static data of the Common Knowledge Base. In addition, this Java application will include in the full prototype the corresponding functionality to administrate the users of the FOKSai system, including the definition of users and user groups, and the definition of rights for each user group, regarding what could be accessed, modified and/or deleted in the system.

AmI Information Processing Module

The main function of the AmI Information Processing Module (AmI IPM) in the early prototype is to map the input XML data to the Common Knowledge Base (CKB). AmI IPM does not have a Graphical User Interface (GUI) in the early prototype, just a command line input, where the input file name can be given, and text based output, to where it writes certain messages to let the user or tester know what is going on during the tests.

This module has been developed using the 1.4 version of J2EE SDK. For the easier development the Eclipse Project software suite was used as a GUI to help the work. For the parsing of XML files the SAX XML Parser was integrated into this module. The early prototype of the AmI IPM - as a standalone Java application - was tested with the simulated data of the business cases. The input data was written into XML files according to the defined structure in the schema file.

Diagnostic Engine

The Diagnostic Engine (DE) provides interactive problem solving assistance to the users, using the structured method of Case-Based Reasoning (CBR) to rapidly get the required problem solving information. This module has been implemented in C++ using the function library of the ReCall [22] tool, and connected to FOKSai system through a CORBA interface. The user uses DE's functionality and accesses its results through the Product Support Module. The testing of this module was done through the assessment of the Product Support Module, as described later in this document and its functionalities are fully hidden from the system user.

Product Support Module

The Product Support Module (PSM) is the central point of interaction between the user and the FOKSai system, i.e. it contains the Graphical User Interface. In addition, this module will include information, documents and knowledge relevant for product and customer support (e.g. information on new product models, new services, advises how to apply products and services in order to avoid problems etc.). The early prototype of the Product Support Module is implemented in Enterprise Java Beans (J2EE1.4), and available through a Java Graphical User Interface only.

Knowledge Analysis Module

The Knowledge Analysis Module (KAM) is comprised of three main functionalities:

1 *Statistical Analysis*: FOKSai users can create Pareto charts of the most common problems by type and severity using the utility. Users also can list all those charted problems using a special utility. This tool is intended to be used for problem identification.

2 *Database Query Tool:* It allows FOKSai users to perform flexible database queries simply selecting the predefined items available. FOKSai administrator can create and store SQL queries in order to allow other users to use them. Administrator can easily create specific queries for CKB Maintenance according the necessities.

3 *Knowledge Analysis:* FOKSai users can use a forum in order to provide statistical analysis reports and send feedback to design staff, create and upload maintenance reports, receive online technical support, obtain quick fixes for current problems or know everything about new developments on products.

The Knowledge Analysis Module has been developed following the Sun J2EE 1.4 standard. It is comprised of Servlets, JSP and EJB and it has been tested in JBoss 4.0.2 Java Application Server. The module is connected to the PSM and the CKB and is accessible using a web browser. At full prototype stage, it will also be accessible through the PSM. In essence, this module allows the users to perform Statistical Analysis of the problems stored in the CKB, lists the problems selected, any table of the CKB for maintenance purposes and provide feedback to FOKSai users via a Forum tool. KAM module uses SSL and form-based authentication, which allows the users to search among all forum elements.

Conclusions

The note gives a bird-eye view on the emerging options offered by distributed intelligence tools to foster competitiveness in the supply chains of industrialised countries. The ICT tools are recognised driving support, basically, on their ability of providing information-intensive aids, in parallel to the traditional trading of manufactured goods. This has falls-off on the value build-up, enhancing direct intangible provisions, and opening indirect opportunities with *product-service* deliveries, on condition that suitable ICT tools are implemented. The new business paradigms are strictly grounded on the availability of the full transparency over the product lifecycle, with profit of the manufacturers (according to economy of scope rules), of the users (for reliable conformance-to-specification management), and the third parties (for better eco-consistency compliance). The prospected analysis moves from the connections that link Product Lifecycle Management, PLM, and Service Engineering, SE, to show how these are the information prerequisites to aim at Condition Monitoring Maintenance, CMM, set-ups, built as

Knowledge Based Systems, to provide the Ambient Intelligence, AmI, consistent with the new business paradigms.

The discussion is restricted to the domain of the ICT co-operative infrastructures, supporting products-services by networked organisations, say, clusters of enterprises that collaborated for any given delivering, with benefit of the customers having a unified responsible over the life-long exploitation of the purchased goods. This certainly does not means that the manufacturer will keep in charge of the whole activity (even if this seems to be prospected by the EU rules, on the suppliers responsibilities), rather than proper out-sourcing could establish, on condition that appropriate PLM/SE tools are provided together with every *extended* artefacts, to make operative the supporting *extended* enterprises.

Acknowledgements

This research work was partially supported by ERCIM (European Research Consortium for Informatics and Mathematics), by the European Union in the frame of FOKSai project and by the Hungarian State Eötvös Scholarship.

References

[1] François B. Vernadat "Enterprise Modeling and Integration: principles and applications" Chapman&Hall, pp. 1-513, 1996

[2] Marco Garetti "PLM: a new business model to foster product innovation", In proceedings at International IMS forum 2004: Global Challenges in Manufacturing, Vol. 2, pp. 917-924, Villa-Erba-Cernobbio, Italy, May 17-19, 2004

[3] Rinaldo C. Michelini, Georg L. Kovacs "Information infrastructures and Sustainability", In book: Emerging Solutions for Future Manufacturing Systems, Springer, pp. 347-356, 2004

[4] F. Ameri, D. Dutta "Product Lifecycle Management Needs, Concepts and Components", Technical Report, http://www.plmdc.engin.umich.edu/, pp. 1-3

[5] V. Liestmann, G. Gudergan, C. Gill "Architecture for Service Engineering – The Design and Development of Industrial Services", In proceedings at International IMS forum 2004: Global Challenges in Manufacturing, Vol. 1, pp. 20-25 249-256, Villa-Erba-Cernobbio, Italy, May 17-19, 2004

[6] I. Rasovska, B. Chebel-Morello, N. Zerhouni "A Conceptual Model of Maintenance Process in Unified Modeling Language" In proceedings at 11th IFAC Symposium on Information Control Problems in Manufacturing (INCOM 2004), pp. 43-48, Salvador, Brasil, April 5-7, 2004

[7] A. Arnaiz, R. Arana, I. Maurtua, L. Susperregi "Maintenance: future technologies" In proceedings at International IMS forum 2004: Global Challenges in Manufacturing, Vol. 1, pp. 300-307, Villa-Erba-Cernobbio, Italy, May 17-19, 2004

[8] Sandra Gross, Elgar Fleisch, "Maintenance Improvement by Unique Product Information Enabled by Ubiquitous Computing", In: 11th IFAC Symposium on Information Control Problems in Manufacturing, pp. 65-70, Salvador, Brasil, April 5-7, 2004

[9] Jay Lee, Hai Qiu, Jun Ni, Dragan Djurdjanovic, "Infotronics Technologies and Predective Tools for Next-generation Maintenance Systems", In proceedings at 11th IFAC Symposium on Information Control Problems in Manufacturing (INCOM 2004), pp. 85-90, Salvador, Brasil, April 5-7, 2004

[10] Projetech company, http://www.projetech.com

[11] G. Riva, F. Vatalaro, F. Davide and M. Alcañiz, "Ambient Intelligence. The Evolution of Technology, Comm. and Cognition Towards the Future of Human-Computer Interaction." IOS Press, pp. 1-320, 2005

[12] Marcus Bengtsson, "Condition Based Maintenance System Technology – Where is Development Heading?" In proceedings of the 17th European Maintenance Congress, May 11-13, 2004, AMS (Spanish Maintenance Society), Barcelona, Spain, B-19.580-2004

[13] Constantine Stephanidis: Ambient Intelligence in the Context of Universal Access, ERCIM News, No. 47, October 2001. pp. 10-11

[14] Mish, F. C. and Morse, J. M. (eds.) (1999) Merriam-Webster's Collegiate Dictionary, tenth edition (Springfield: Merriam-Webster, Inc.), p. 36

[15] ISTAG, IST Advisory Group, Advisory Group to the European Community's Information Society Technology Program http://www.cordis.lu/ist/istag.htm

[16] Gupta, M. (2003) "Ambient Intelligence - unobtrusive technology for the information society". *Pressbox.co.uk*, June 17 http://www.pressbox.co.uk/Detailed/7625.html

[17] Weiser, M. (1991) "The Computer for the 21st Century". *Scientific American*, Vol. 265, No. 3, September, 94-104 http://www.ubiq.com/hypertext/weiser/SciAmDraft3.html

[18] Abowd, G.D. (2004) Investigating Research Issues in Ubiquitous Computing: The Capture, Integration, and Access Problem

http://www.cc.gatech.edu/fce/c2000/pubs/nsf97/summary.html

[19] Van Loenen, E. J. (2003) "Ambient intelligence: Philips' vision". Presentation at *ITEA 2003*, Oulu, Finland, January 10 http://www.vtt.fi/ele/new/ambience/evert_van_loenen.ppt (8.5 MB Powerpoint slide presentation)

[20] Ailisto, H., Kotila, A. and Strömmer, E. (2003) "Ubicom applications and technologies". Presentation slides from *ITEA 2003*, Oulu, Finland http://www.vtt.fi/ict/publications/ailisto_et_al_030821.pdf

[21] SME Focussed KM System to support extended product in Ambient Intelligence domain, 6th FP, COOP-CT-2003-508637, Annex I – Description of Work, 2003

[22] http://www.alice-soft.com/html/prod_recall.htm

[23] The extended products paradigm. An introduction. Jansson, Kim; Thoben, Klaus-Dieter DIISM2002 - The 5th International Conference on Design of Information Infrastructure Systems for Manufacturing 2002, Osaka, Japan

[24] EXPIDE Project http://www.expide.org (30.4.2003)

Stability and Sensitivity Analysis of Fuzzy Control Systems. Mechatronics Applications

Radu-Emil Precup, Stefan Preitl

"Politehnica" University of Timisoara, Dept. of Automation and Appl. Inform.
Bd. V. Parvan 2, RO-300223 Timisoara, Romania
Phone: +40-256-4032-29, -30, -24, -26, Fax: +40-256-403214
E-mail: radu.precup@aut.upt.ro, stefan.preitl@aut.upt.ro

Abstract: The development of fuzzy control systems is usually performed by heuristic means, incorporating human skills, the drawback being in the lack of general-purpose development methods. A major problem, which follows from this development, is the analysis of the structural properties of the control system, such as stability, controllability and robustness. Here comes the first goal of the paper, to present a stability analysis method dedicated to fuzzy control systems with mechatronics applications based on the use of Popov's hyperstability theory. The second goal of this paper is to perform the sensitivity analysis of fuzzy control systems with respect to the parametric variations of the controlled plant for a class of servo-systems used in mechatronics applications based on the construction of sensitivity models. The stability and sensitivity analysis methods provide useful information to the development of fuzzy control systems. The case studies concerning fuzzy controlled servo-systems, accompanied by digital simulation results and real-time experimental results, validate the presented methods.

Keywords: Mamdani fuzzy controllers, stability analysis, sensitivity analysis, mechatronics, servo-systems.

1 Introduction

The development of fuzzy control systems (FCSs) is usually performed by heuristic means, due to the lack of general development methods applicable to large categories of systems. A major problem, which follows from the heuristic method of development, is the analysis of the structural properties of the control systems including the stability analysis and the sensitivity analysis with respect to the parametric variations of the controlled plant. In case of mechatronics applications focussed on servo-systems the analysis of these properties becomes more important due to the very good steady-state and dynamic performance they must ensure. Therefore, the paper aims a twofold goal. Firstly, it presents one stability analysis method dedicated to FCSs applied to servo-systems with mechatronics applications, involving the use of Popov's hyperstability theory. Secondly, the paper performs the sensitivity analysis of FCSs with respect to the parametric variations of the controlled plant (CP) for a class of servo-systems based on the construction of sensitivity models. The considered FCSs contain Mamdani fuzzy controllers with singleton consequents, seen as type-II fuzzy systems [1, 2].

The stability analysis of an FCS is justified because only a stable FCS can ensure the functionality of the plant and, furthermore, the disturbance reduction, guarantee desired steady states, and reduce the risk of implementing the fuzzy controller (FC). The main approaches in stability analysis of FCSs with Mamdani fuzzy controllers concern: the state-space approach based on a linearized model of the nonlinear system [3, 4], Popov's hyperstability theory [5, 6], Lyapunov's direct method [2, 7], the circle criterion [7, 8], the harmonic balance method [8, 9] referred to also as the describing function method [10, 11], the passivity approach [12], etc.

The sensitivity analysis of the FCSs with respect to the parametric variations of the CP is necessary because the behaviour of these systems is generally reported as "robust" or "insensitive" without offering systematic analysis tools. The sensitivity analysis performed in the paper is based on the idea of approximate equivalence, in certain conditions, between FCSs and linear ones. This is fully justified because of two reasons. The first reason is related with the controller part of the FCS, where the approximately equivalence between linear and fuzzy controllers is generally acknowledged [13, 14]. The second reason is related with the plant part of the FCS. The support for using an FC developed to control a plant having a linear or linearized model is in the fact that this plant model can be considered as a simplified model of a relatively complex model of the CP having nonlinearities or variable parameters or being placed at the lower level of large-scale systems. This is the case of servo-systems in mechatronics applications. Although the plant is nonlinear, it can be linearized in the vicinity of a set of operating points or of a trajectory. The plant model could be also uncertain or not well defined. The FC, as essentially nonlinear element, can compensate for the model uncertainties, nonlinearities and parametric variations of the CP. Fuzzy control must not be seen as a goal in itself, but sometimes the only way to initially approach the control of complex plants.

This paper is organized as follows. It will be treated in the following Section the stability analysis method based on the use of Popov's hyperstability theory. The exemplification of the method is done by a case study regarding the FC development to control an electro-hydraulic servo-system. Then, in Section 3 an approach to the sensitivity analysis of the FCSs for a class of servo-systems with respect to the parametric variations of the CP is presented. This approach is illustrated by a case study regarding the fuzzy control of a nonlinear servo-system. Digital simulation results and real-time experimental results validate the presented approaches. The conclusions are drawn in the end of the paper.

2 Stability Analysis Method Based on Popov's Hyperstability Theory

The SAM based on Popov's hyperstability theory is applied to FCSs to control SISO plants when employing PI-fuzzy controllers (PI-FCs). The structure of the considered FCS is a conventional one, presented in Fig. 1 (a), where: r – the reference input, y – the controlled output, e – the control error, u – the control signal, d_1, d_2, d_3 – the disturbance inputs.

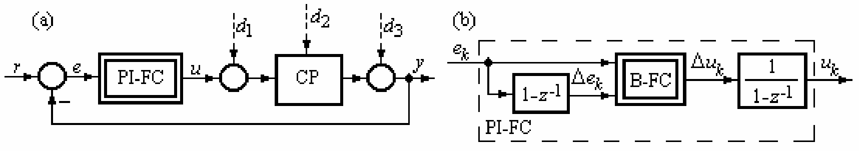


Figure 1
Structure of FCS (a) and of PI-FC (b)

The CP includes the actuator and the measuring devices. The application of an FC, when conditions for linear operating regimes of the plant are validated, determines the FCS to be considered as a Lure-Postnikov type nonlinear control system (for example, see [15]). The PI-FC represents a discrete-time FC with dynamics, introduced by the numerical differentiation of the control error e_k expressed as the increment of control error, $\Delta e_k = e_k - e_{k\text{-}1}$, and by the numerical integration of the increment of control signal, Δu_k. The structure of the considered PI-FC is illustrated in Fig. 1 (b), where B-FC represents the basic fuzzy controller, without dynamics.

The block B-FC is a nonlinear two inputs-single output (TISO) system, which includes among its nonlinearities the scaling of inputs and output as part of its fuzzification module. The fuzzification is solved in terms of the regularly distributed (here) input and output membership functions illustrated in Fig. 2. Other distributions of the membership functions can modify in a desired way the controller nonlinearities.

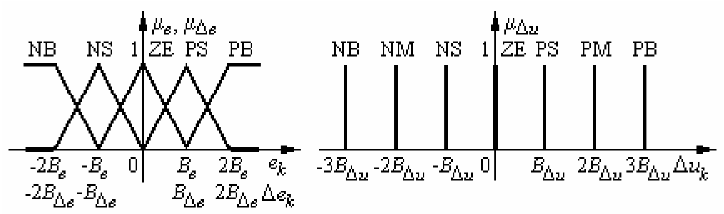


Figure 2
Membership functions of input and output linguistic variables of B-FC

The inference engine in B-FC employs Mamdani's MAX-MIN compositional rule of inference assisted by the rule base presented in Table 1, and the center of gravity method for singletons is used for defuzzification.

Table 1

Decision table of B-FC

$\Delta e_k \setminus e_k$	NB	NS	ZE	PS	PB
PB	ZE	PS	PM	PB	PB
PS	NS	ZE	PS	PM	PB
ZE	NM	NS	ZE	PS	PM
NS	NB	NM	NS	ZE	PS
NB	NB	NB	NM	NS	ZE

To develop the PI-FC the beginning is in the expression of the discrete-time equation of a digital PI controller in its incremental version:

$$\Delta u_k = K_P \cdot \Delta e_k + K_I \cdot e_k = K_P (\Delta e_k + \alpha \cdot e_k), \quad (1)$$

where k is the index of the current sampling interval.

In case of a quasi-continuous digital PI controller the parameters K_P, K_I and α can be calculated as functions of the parameters k_C (gain) and T_i (integral time constant) of a basic original continuous-time PI controller having the transfer function $H_C(s)$:

$$H_C(s) = [k_C/(sT_i)](1 + sT_i), \quad (2)$$

and the connections between $\{K_P, K_I, \alpha\}$ and $\{k_C, T_i\}$ have the following form in the case of using Tustin's discretization method:

$$K_P = k_C[1 - T_s/(2T_i)],\ K_I = k_C T_s / T_i,\ \alpha = K_I / K_P = 2T_s/(2T_i - T_s), \quad (3)$$

with T_s – the sampling period chosen in accordance with the requirements of quasi-continuous digital control.

The design relations for the PI-FC are obtained by the application of the modal equivalence principle [16] transformed into (4) in this case:

$$B_{\Delta e} = \alpha B_e,\ B_{\Delta u} = K_I B_e, \quad (4)$$

where the free parameter B_e represents designer's option. Using the experience in controlling the plant one can choose the value of this parameter, but firstly it must be chosen to ensure the aim of a stable FCS.

The CP is supposed to be characterized by the following n-th order discrete-time SISO linear time-invariant state mathematical model (MM) including the zero-order hold:

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{A}\cdot\mathbf{x}_k + \mathbf{b}\cdot u_k \\ y_k &= \mathbf{c}^T\cdot\mathbf{x}_k \end{aligned}, \tag{5}$$

where: u_k – the control signal; y_k – the controlled output; $\mathbf{x}_k$ – the state vector, dim $\mathbf{x}_k$ = $(n,1)$; $\mathbf{A}$, $\mathbf{b}$, $\mathbf{c}^T$ - matrices with the dimensions: dim $\mathbf{A} = (n, n)$, dim $\mathbf{b} = (n, 1)$, dim $\mathbf{c}^T = (1, n)$.

To derive the stability analysis method it is necessary to transform the initial FCS structure into a multivariable one because the block B-FC in Fig.X.5 is a TISO system. This modified FCS structure is illustrated in Fig. 3 (a), where the dynamics of the fuzzy controller (its linearized part) is transferred to the plant (CP) resulting in the extended controlled plant (ECP, a linear one). The vectors in Fig. 3 (a) represent: $\mathbf{r}_k$ – the reference input vector, $\mathbf{e}_k$ – the control error vector, $\mathbf{y}_k$ – the controlled output vector, $\mathbf{u}_k$ – the control signal vector. For the general use (in the continuous time, too) the index k may be omitted, and these vectors are defined as follows:

$$\mathbf{r}_k = [r_k \quad \Delta r_k]^T,\ \mathbf{e}_k = [e_k \quad \Delta e_k]^T,\ \mathbf{y}_k = [y_k \quad \Delta y_k]^T, \tag{6}$$

where $\Delta v_k = v_k - v_{k-1}$ stands generally for the increment of the variable v_k.

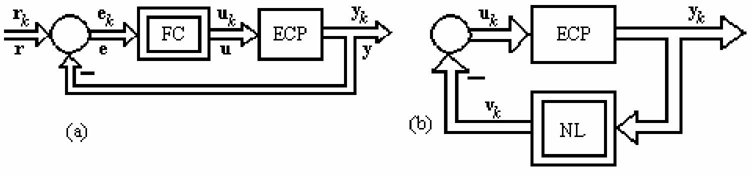


Figure 3

Modified structure of FCS (a) and structure used in stability analysis (b)

In relation with Fig. 3 (a), the block FC is characterized by the nonlinear input-output static map $\mathbf{F}$:

$$\mathbf{F}: R^2 \to R^2,\ \mathbf{F}(\mathbf{e}_k) = [f(\mathbf{e}_k),\ 0]^T \ , \tag{7}$$

where f ($f : R^2 \to R$) is the input-output static map of the nonlinear TISO system B-FC in Fig. 1.

As it can be observed in (6), all variables in the FCS structure (in Fig. 3 (a)) have two components. This requires the introduction of a fictitious control signal, supplementary to the outputs of the block B-FC, for obtaining an equal number of inputs and outputs as required by the hyperstability theory in the multivariable case.

Generally speaking, the structure involved in the stability analysis of an unforced nonlinear control system ($\mathbf{r}_k$ = $\mathbf{0}$ and the disturbance inputs are also zero) is presented in Fig. 3 (b). The block NL in Fig. 3 (b) represents a static nonlinearity due to the nonlinear part without dynamics of the block FC in Fig. 3 (a). The connections between the variables of the control system structures in Fig. 3 are:

$$\mathbf{v}_k = -\mathbf{u}_k = -\mathbf{F}(\mathbf{e}_k), \ \mathbf{y}_k = -\mathbf{e}_k, \tag{8}$$

with the second component of **F** being always zero to neglect the effect of the fictitious control signal.

The MM of the ECP can be derived by firstly defining the additional state variables $\{x_{uk}, x_{yk}\}$ according to Fig. 4. Then, the extended state vector $\mathbf{x}_k^E$ and the control signal vector $\mathbf{u}_k^E$ can be expressed in terms of (9):

$$\mathbf{x}_k^E = \begin{bmatrix}\mathbf{x}_k^T & x_{uk} & x_{yk}\end{bmatrix}^T, \ \mathbf{u}_k^E = \begin{bmatrix}\Delta u_k & \Delta u_{fk}\end{bmatrix}^T, \tag{9}$$

where Δu_{fk} stands for the fictitious increment of control signal.

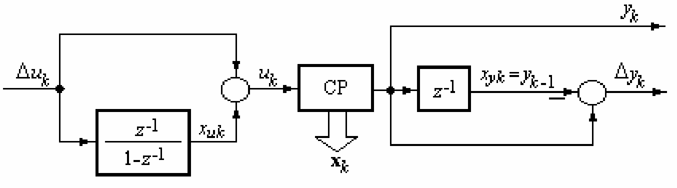


Figure 4
Structure of ECP block

Using the structure presented in Fig. 4, the (n+2)-th order discrete-time state MM of the ECP becomes (10):

$$\begin{aligned}\mathbf{x}_{k+1}^E &= \mathbf{A}^E \cdot \mathbf{x}_k^E + \mathbf{B}^E \cdot \mathbf{u}_k^E \\ \mathbf{y}_k^E &= \mathbf{C}^E \cdot \mathbf{x}_k^E\end{aligned}, \tag{10}$$

with the matrices $\mathbf{A}^E$ (dim $\mathbf{A}^E = (n+2, n+2)$), $\mathbf{B}^E$ (dim $\mathbf{B}^E = (n+2, 2)$) and $\mathbf{C}^E$ (dim $\mathbf{C}^E = (2, n+2)$) expressed as follows:

$$\mathbf{A}^E = \begin{bmatrix}\mathbf{A} & \mathbf{b} & \mathbf{0} \\ \mathbf{0}^T & 1 & 0 \\ \mathbf{c}^T & 0 & 0\end{bmatrix}, \ \mathbf{B}^E = \begin{bmatrix}\mathbf{b} & \mathbf{1} \\ 1 & 1 \\ 0 & 1\end{bmatrix}, \ \mathbf{C}^E = \begin{bmatrix}\mathbf{c}^T & 0 & 0 \\ \mathbf{c}^T & 0 & -1\end{bmatrix}. \tag{11}$$

Then, the second equations in (8) and (10) can be transformed into the following equivalent expressions:

$$\mathbf{e}_k = -\mathbf{C}^E \cdot \mathbf{x}_k, \ \mathbf{x}_k = \mathbf{C}^b \cdot \mathbf{e}_k, \tag{12}$$

where the matrix $\mathbf{C}^b$, dim $\mathbf{C}^b = (n+2, 2)$, can be calculated relatively easy as function of $\mathbf{C}^E$.

The main part of the proposed stability analysis method can be stated in terms of the following theorem giving sufficient stability conditions.

Theorem 1. The nonlinear system, with the structure presented in Fig. 3 (b) and the MM of its linear part (10), is globally asymptotically stable if:

- the three matrices **P** (positive definite, dim **P** = (n+2,n+2)), **L** (regular, dim **L** = (n+2,n+2)) and **V** (any, dim **V** = (n+2,2)) fulfill the requirements (13):

$$\begin{aligned}&(\mathbf{A}^E)^T \cdot \mathbf{P} \cdot \mathbf{A}^E = -\mathbf{L} \cdot \mathbf{L}^T\\ &\mathbf{C}^E - (\mathbf{B}^E)^T \cdot \mathbf{P} \cdot \mathbf{A}^E = \mathbf{V}^T \cdot \mathbf{L}^T ,\\ &-(\mathbf{B}^E)^T \cdot \mathbf{P} \cdot \mathbf{B}^E = \mathbf{V}^T \cdot \mathbf{V}\end{aligned} \tag{13}$$

- introducing the matrices **M** (dim **M** = (2,2)), **N** (dim **N** = (2,2)) and **R** (dim **R** = (2,2)) defined in terms of (14):

$$\begin{aligned}&\mathbf{M} = (\mathbf{C}^b)^T \cdot (\mathbf{L} \cdot \mathbf{L}^T - \mathbf{P}) \cdot \mathbf{C}^b\\ &\mathbf{N} = (\mathbf{C}^b)^T \cdot [\mathbf{L} \cdot \mathbf{V} - (\mathbf{A}^E)^T \cdot \mathbf{P} \cdot \mathbf{B}^E - 2(\mathbf{C}^E)^T],\\ &\mathbf{R} = \mathbf{V}^T \cdot \mathbf{V}\end{aligned} \tag{14}$$

the Popov-type inequality (15) holds for any value of the control error e_k:

$$f(\mathbf{e}_k) \cdot \mathbf{n}^T \cdot \mathbf{e}_k + (\mathbf{e}_k)^T \cdot \mathbf{M} \cdot \mathbf{e}_k \geq 0 , \tag{15}$$

where **n** represents the first column in **N**.

The proof of Theorem 1, based on the Kalman-Szegö lemma [17] and on processing the Popov sum, is presented in [18] for the PI-FCs with prediction.

By taking into account these aspects, the stability analysis method dedicated to FCSs with PI-FCs consists of the following steps:

- step (a): express the MM of the CP, choose the sampling period T_s and calculate the discrete-time state-space MM of the CP with the zero-order hold, (5),
- step (b): derive the discrete-time state-space mathematical model of the ECP, (10),
- step (c): compute the matrix $\mathbf{C}^b$ in terms of (12),
- step (d): solve the system of equations (13), with the solutions **P**, **L** and **V**, and calculate the matrices **M**, **N** and **R** in (14),
- step (e): set the value of the free parameter $B_e > 0$ of the PI-FC, and tune the other parameters of the PI-FC in terms of (4),
- step (f): check the stability condition (15) for any values of PI-FC inputs in operating regimes considered to be significant for FCS behaviour.

To test the presented stability analysis method it is considered the CP of an electro-hydraulic servo-system (EHS) used as actuator in mechatronics applications, with the structure presented in Fig. 5 (a) [19], where: NL 1 … NL 5 – nonlinearities, EHC – electro-hydraulic converter, SVD – slide-valve distributor, MSM – main servo-motor, M 1 and M 2 – measuring devices; u – control signal, y – controlled output; x_1 and x_2 – state variables; x_{1M} and x_{2M} – measured state variables, $u_l = 10$ V, $g_0 = 0.0625$ mm/V, $\varepsilon_2 = 0.02$ mm, $\varepsilon_4 = 0.2$ mm, $x_{1l} = 21.8$ mm, $y_l = 210$ mm, $T_{i1} = 0.001872$ sec, $T_{i2} = 0.0756$ sec, $k_{M1} = 0.2$ V/mm, $k_{M2} = 0.032$ V/mm. To obtain a relatively simple FC, the CP is represented here by the stabilized electro-hydraulic servo-system (SEHS), and the FCS structure is presented in Fig. 5 (b).

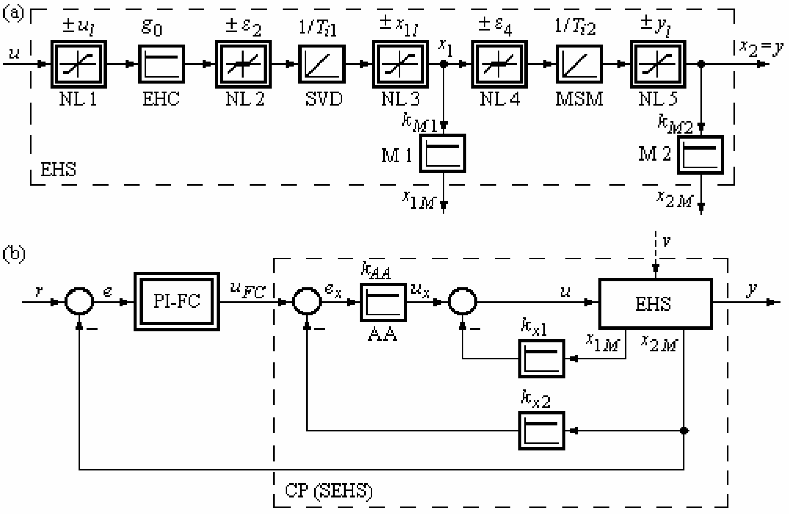


Figure 5
Structure of EHS as CP (a) and structure of FCS (b)

The SEHS represents itself a state feedback control system, with AA – adder amplifier, and k_{x1}, k_{x2}, k_{AA} – parameters of the state feedback controller. By omitting the nonlinearities of the EHS, imposing the double pole of the SEHS in −1, the pole placement method leads to the transfer function of the SEHS block, $H_{CP}(s)$:

$$H_{CP}(s) = \frac{1/k_{x2}}{1+[k_{M1}T_{i2}k_{x1}/(k_{AA}k_{M2}k_{x2})]s+[T_{i1}T_{i2}/(g_0k_{AA}k_{M2})]s^2} = \frac{1}{(1+s)^2}, \tag{16}$$

obtained for $k_{x1} = 0.2997$, $k_{x2} = 1$, $k_{AA} = 0.0708$.

The steps of the stability analysis method have been proceeded, but only the values of $\mathbf{M}$ and $\mathbf{n}^T$ are presented because these two matrices appear in the stability condition (15), tested by digital simulation:

$$\mathbf{M} = \begin{bmatrix} 2.0071 & 0 \\ 0 & 0 \end{bmatrix}, \ \mathbf{n}^T = \begin{bmatrix} 0.9855 & 0 \end{bmatrix}, \tag{17}$$

and the parameters of the PI-FC ensuring the stability of the FCS have been tuned as: $B_e = 0.3$, $B_{\Delta e} = 0.0076$, $B_{\Delta u} = 0.0203$. To verify the stability of all FCS the dynamic behaviour of the free control system was simulated, when the system started from two different, arbitrarily chosen, initial states, obtained by feeding a d_3=–0.5 and a d_3=–1.5 disturbance input according to Fig. 1 (a). The FCS behaviour is presented in Fig. 6, and it illustrates that the FCS analyzed by using the presented SAM is stable.

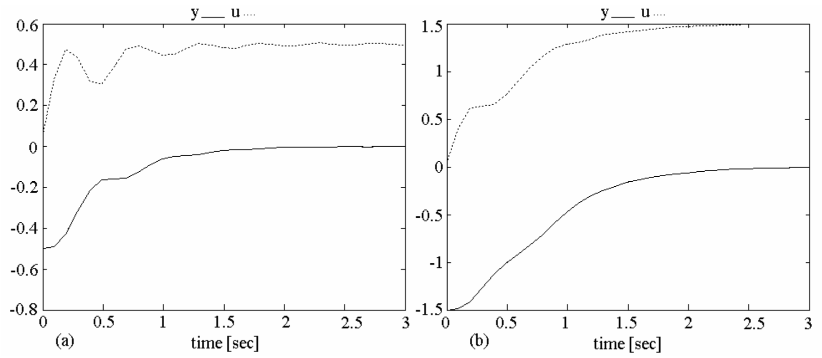


Figure 6

FCS behaviour for d_3 = – 0.5 (a) and d_3 = – 1.5 (b)

3 Sensitivity Analysis of a Class of Fuzzy Control Systems. Case Study

Let the considered control system structure be a conventional one, presented in Fig. 7 (a), where: C – controller, CP – controlled plant, RF – the reference filter, r – reference input, $\tilde{r}$ – filtered reference input, e – control error, u – control signal, y – controlled output, d_1, d_2, d_3, d_4 – disturbance inputs. Depending on the place of feeding the disturbance inputs to the CP and on the CP structure, the accepted types of disturbance inputs $\{d_1, d_2, d_3, d_4\}$ are defined in terms of Fig. 7 (b).

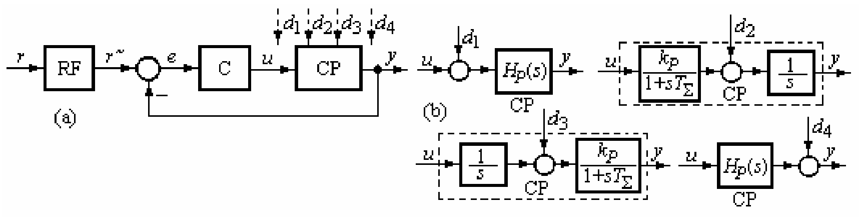


Figure 7

Control system structure (a) and disturbance inputs types (b)

The class of plants with the simplified structure illustrated in Fig. 7 (a), is considered to be linearized around a steady-state operating point and characterized by the transfer function $H_P(s)$:

$$H_P(s) = k_P /[s(1 + T_\Sigma s)]\,, \tag{18}$$

with k_P – gain and T_Σ – small time constant or sum of all parasitic time constants, belongs to a class of integral-type systems with variable parameters in case of servo-systems with mechatronics applications.

For these plants, it is recommended the use of linear PI controllers having the transfer function $H_C(s)$ in (2) with $k_C = k_c T_i$. Based on the Extended Symmetrical Optimum (ESO) method [19], the parameters of the controller {k_c (or k_C) – controller gain, T_i – integral time constant} are tuned in terms of (19) guaranteeing the desired control system performance by means of a single design parameter, β:

$$k_c = 1/(\beta^{3/2} k_P T_\Sigma^2),\ \ T_i = \beta T_\Sigma,\ k_C = 1/(\beta^{1/2} k_P T_\Sigma)\,. \tag{19}$$

The PI controllers can be tuned also in terms of the Iterative Feedback Tuning (IFT) method, representing a data-based design method where the update of the controller parameters is done through an iterative procedure. IFT is a gradient-based approach, based on input-output data recorded from the closed-loop system. The control system performance indices are specified by the proper expression of a criterion function. Optimizing such functions usually requires iterative gradient-based minimization, and this can be a complicated function of the plant and of the disturbances dynamics. The key property of IFT is that the closed-loop experimental data are used to calculate the estimated gradient of the criterion function. Several experiments are performed at each iteration and, based upon the input-output data collected from the system, the updated controller parameters are obtained. Theoretical and practical applications of IFT have been reported in [20, 21, 22, 23]. Using the IFT as a design step in the development of fuzzy controllers can result in efficient development techniques for fuzzy controller with dynamics.

According to [24], the variation of CP parameters ({k_P, T_Σ} for the considered class of CPs) due to the change of the steady-state operating points or to other conditions leads to additional motion (of the control systems). This motion is usually undesirable under uncontrollable parametric variations. Therefore, to alleviate the effects of parametric disturbances it is necessary to perform the sensitivity analysis with respect to the parametric variations of the CP.

It is generally accepted that FCs ensure control system performance enhancement with respect to the modifications of the reference input, of the load disturbance inputs or to parametric variations. This justifies the research efforts focused on the systematic analysis of FCSs behaviour with respect to parametric variations and the need to perform the sensitivity analysis with this respect that enables to derive sensitivity models for the FCs and for the overall FCSs.

The sensitivity models enable the sensitivity analysis of the FCSs accepted, as mentioned in Section 1, to be approximately equivalent with the linear control systems. This justifies the approach to be presented in the sequel, that the sensitivity models of the FCSs are approximately equivalent to the sensitivity models of the linear ones. Therefore, it is necessary to obtain firstly the sensitivity models of the linear control system in Fig. 7 (a).

Defining the state sensitivity functions {λ$_1$, λ$_2$, λ$_3$} and the output sensitivity function, σ, there have been derived several sensitivity models, four of them being presented as follows:

- with respect to the variation of k_P, the step modification of r, and $d_3(t) = 0$:

$$\begin{aligned}
&\dot{\lambda}_1(t) = \lambda_2(t),\\
&\dot{\lambda}_2(t) = -[1/(\beta^{1/2}T_{\Sigma 0}^2)]\lambda_1(t) - (1/T_{\Sigma 0})\lambda_2(t) + [1/(\beta^{1/2}T_{\Sigma 0}^2)]\lambda_3(t) -\\
&\qquad - [1/(\beta^{1/2}k_{P0}T_{\Sigma 0}^2)]x_{10}(t) + [1/(\beta^{1/2}k_{P0}T_{\Sigma 0}^2)]x_{30}(t) +\\
&\qquad + [1/(\beta^{1/2}k_{P0}T_{\Sigma 0}^2)]r_0(t),\\
&\dot{\lambda}_3(t) = -[1/(\beta T_{\Sigma 0})]\lambda_1(t),\\
&\sigma(t) = \lambda_1(t),
\end{aligned} \tag{20}$$

- with respect to the variation of T_Σ, the step modification of r, and $d_3(t) = 0$:

$$\begin{aligned}
&\dot{\lambda}_1(t) = \lambda_2(t),\\
&\dot{\lambda}_2(t) = -[1/(\beta^{1/2}T_{\Sigma 0}^2)]\lambda_1(t) - (1/T_{\Sigma 0})\lambda_2(t) + [1/(\beta^{1/2}T_{\Sigma 0}^2)]\lambda_3(t) +\\
&\qquad + [1/(\beta^{1/2}T_{\Sigma 0}^3)]x_{10}(t) + (1/T_{\Sigma 0}^2)x_{20}(t) - [1/(\beta^{1/2}T_{\Sigma 0}^3)]x_{30}(t) -\\
&\qquad - [1/(\beta^{1/2}T_{\Sigma 0}^3)]r_0(t),\\
&\dot{\lambda}_3(t) = -[1/(\beta T_{\Sigma 0})]\lambda_1(t),\\
&\sigma(t) = \lambda_1(t),
\end{aligned} \tag{21}$$

- with respect to the variation of k_P, the step modification of d_3, and $r(t) = 0$:

$$\begin{aligned}
&\dot{\lambda}_1(t) = -(1/T_{\Sigma 0})\lambda_1(t) + (k_{P0}/T_{\Sigma 0})\lambda_2(t) + (1/T_{\Sigma 0})x_{20}(t) +\\
&\qquad + (1/T_{\Sigma 0})d_{30}(t),\\
&\dot{\lambda}_2(t) = -[1/(\beta^{1/2}k_{P0}T_{\Sigma 0})]\lambda_1(t) + [1/(\beta^{1/2}k_{P0}T_{\Sigma 0})]\lambda_3(t),\\
&\dot{\lambda}_3(t) = -[1/(\beta T_{\Sigma 0})]\lambda_1(t),\\
&\sigma(t) = \lambda_1(t),
\end{aligned} \tag{22}$$

- with respect to the variation of T_Σ, the step modification of d_3, and $r(t) = 0$:

$$\begin{aligned}
\dot{\lambda}_1(t) &= -(1/T_{\Sigma 0})\lambda_1(t) + (k_{P0}/T_{\Sigma 0})\lambda_2(t) + (1/T_{\Sigma 0}^2)x_{10}(t) - \\
&\quad - (k_{P0}/T_{\Sigma 0}^2)x_{20}(t) - (k_{P0}/T_{\Sigma 0}^2)d_{30}(t), \\
\dot{\lambda}_2(t) &= -[1/(\beta^{1/2}k_{P0}T_{\Sigma 0})]\lambda_1(t) + [1/(\beta^{1/2}k_{P0}T_{\Sigma 0})]\lambda_3(t), \\
\dot{\lambda}_3(t) &= -[1/(\beta T_{\Sigma 0})]\lambda_1(t), \\
\sigma(t) &= \lambda_1(t).
\end{aligned} \tag{23}$$

The behaviours of these four sensitivity models, for the unit step modification of r followed by a unit step modification of d_3 after 250 sec, starting with the initial conditions $\lambda_1(0) = 2$, $\lambda_2(0) = 1$, $\lambda_3(0) = 0$, are shown in Fig. 8.

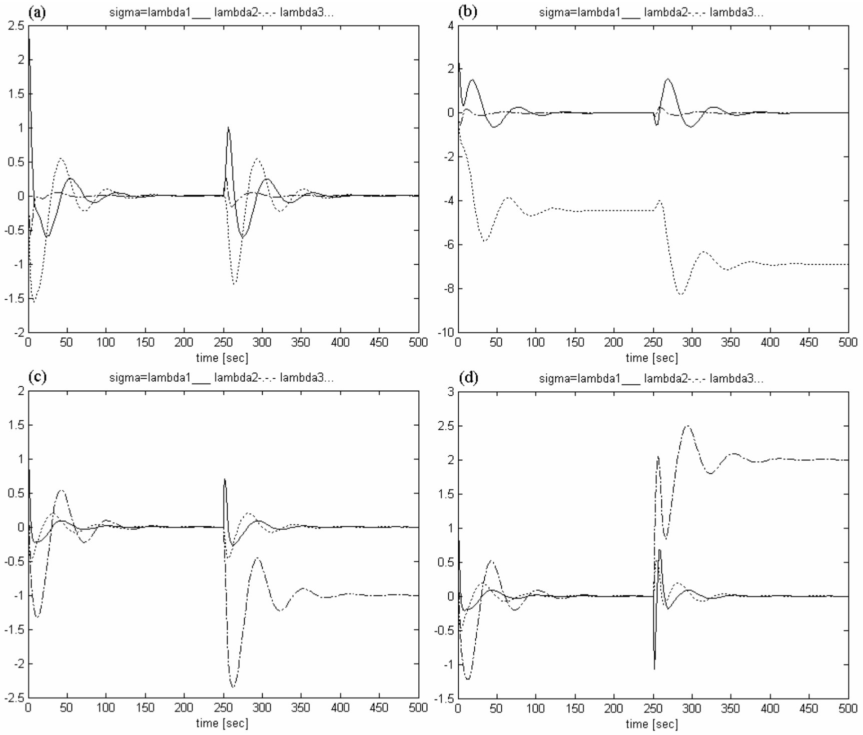


Figure 8

Behaviour of sensitivity models (20) (in (a)), (21) (in (b)), (22) (in (c)) and (23) (in (d))

The PI-FC development method which can be expressed by using the stability and sensitivity analyses presented in this paper is applied in case of a nonlinear laboratory DC drive, AMIRA DR300.

The DC motor is loaded using a current controlled DC generator, mounted on the same shaft, and the drive has built-in analog current controllers for both DC machines with rated speed of 3000 rpm, rated power equal to 30 W, and rated current equal to 2 A. The speed control of the DC motor is digitally implemented using an A/D – D/A data converter card. The speed sensors are a tacho generator and an additional incremental rotary encoder mounted at the free drive-shaft.

The schema of the experimental setup is presented in Fig. 9.

In these conditions, the speed response of the FCS with RF and PI-FC with respect to the modification of the reference input and without load is presented in Fig. 10.

Due to the integral feature of the PI-fuzzy controller structure it is not necessary to present the control system behaviour with respect to the modifications of the load disturbance inputs.

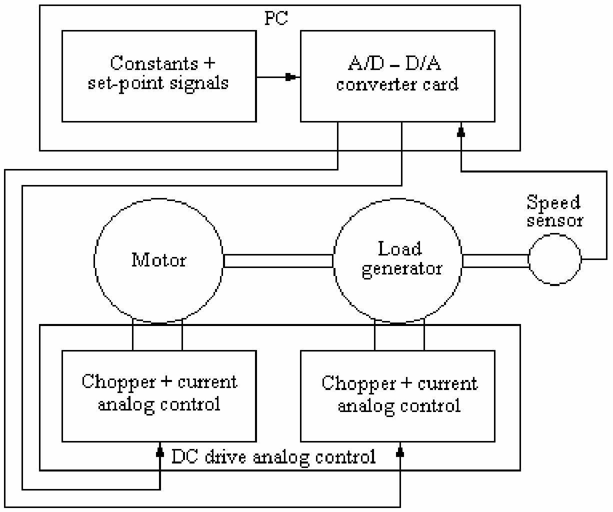

PC
Constants + set-point signals
A/D – D/A converter card
Speed sensor
Motor
Load generator
Chopper + current analog control
Chopper + current analog control
DC drive analog control

Figure 9
Experimental setup schema

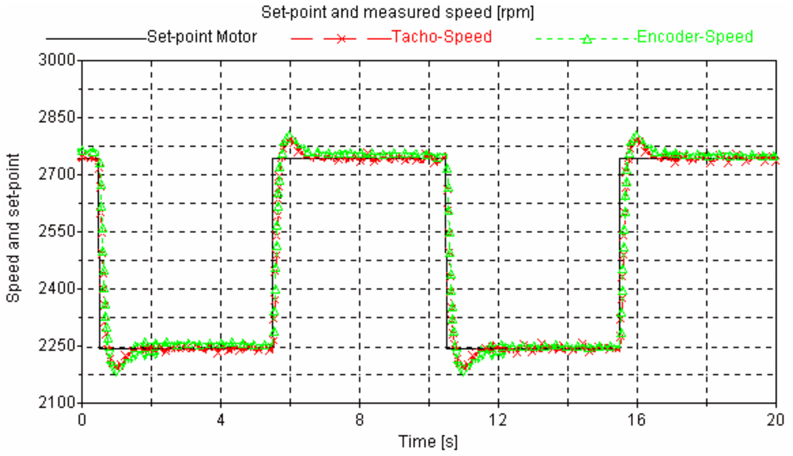


Figure 10

Speed response of FCS without load

Conclusions

The paper presents one stability analysis method and performs the sensitivity analysis of fuzzy control systems with Mamdani fuzzy controllers dedicated to control of servo-systems in mechatronics applications.

The presentation is focused on PI-fuzzy controllers, but it can be applied with no major problems in case of PD- or PID-fuzzy controllers and of complex fuzzy controller structures as well [25, 26, 27].

The methods can be formulated under the form of useful design recommendations for the fuzzy controllers.

The case studies presented in the paper, accompanied by digital simulation results and by experimental results, validate the theoretical approaches.

References

[1] Kóczy, L. T.: Fuzzy If … Then rule models and their transformation into one another, IEEE Transactions on Systems, Man, and Cybernetics - Part A, 1996, Vol. 26, pp. 621-637

[2] Sugeno, M.: On stability of fuzzy systems expressed by fuzzy rules with singleton consequents, IEEE Transactions on Fuzzy Systems, 1999, Vol. 7, pp. 201-224

[3] Aracil, J., A. Ollero, and A. Garcia-Cerezo: Stability indices for the global analysis of expert control systems, IEEE Transactions on Systems, Man, and Cybernetics, 1989, Vol. 19, pp. 998-1007

[4] Precup, R.-E., S. Doboli, and S. Preitl: Stability analysis and development of a class of fuzzy control systems, Engineering Applications of Artificial Intelligence, 2000, Vol. 13, pp. 237-247

[5] Opitz, H.-P.: Fuzzy control and stability criteria, Proceedings of First EUFIT'93 European Congress, Aachen, Germany, 1993, Vol. 1, pp. 130-136

[6] Precup, R.-E. and S. Preitl: Popov-type stability analysis method for fuzzy control systems, Proceedings of Fifth EUFIT'97 European Congress, Aachen, Germany, 1997, Vol. 2, pp. 1306-1310

[7] Passino, K. M. and S. Yurkovich: Fuzzy Control, Addison Wesley Longman, Inc., Menlo Park, CA, 1998

[8] Driankov, D., H. Hellendoorn and M. Reinfrank: An Introduction to Fuzzy Control, Springer-Verlag, Berlin, Heidelberg, New York, 1993

[9] Kiendl, H.: Harmonic balance for fuzzy control systems, Proceedings of First EUFIT'93 European Congress, Aachen, Germany, 1993, Vol. 1, pp. 137-141

[10] Ying, H.: Analytical structure of a two-input two-output fuzzy controller and its relation to PI and multilevel relay controllers, Fuzzy Sets and Systems, 1994, Vol. 63, pp. 21-33

[11] Gordillo, F., J. Aracil and T. Alamo: Determining limit cycles in fuzzy control systems, Proceedings of FUZZ-IEEE'97 Conference, Barcelona, Spain, 1997, pp. 193-198

[12] Calcev, G.: Some remarks on the stability of Mamdani fuzzy control systems, IEEE Transactions on Fuzzy Systems, 1998, Vol. 6, pp. 436-442

[13] Tang, K. L. and R. J. Mulholland: Comparing fuzzy logic with classical controller designs, IEEE Transactions on Systems, Man, and Cybernetics, 1987, Vol. 17, pp. 1085-1087

[14] Moon, B. S.: Equivalence between fuzzy logic controllers and PI controllers for Single Input systems, Fuzzy Sets and Systems, 1995, Vol. 69, pp. 105-113

[15] Hill, D. J. and C. N. Chong: Lyapunov functions of Lure-Postnikov form for structure preserving models of power plants, Automatica, 1989, Vol. 25, pp. 453-460

[16] Galichet, S. and L. Foulloy: Fuzzy controllers: synthesis and equivalences, IEEE Transactions on Fuzzy Systems, 1995, Vol. 3, pp. 140-148

[17] Landau, I. D.: Adaptive Control, Marcel Dekker, Inc., New York, 1979

[18] Precup, R.-E., S. Preitl, and G. Faur: PI predictive fuzzy controllers for electrical drive speed control: methods and software for stable development, Computers in Industry, 2003, Vol. 52, pp. 253-270

[19] Preitl, S. and R.-E. Precup: An extension of tuning relations after Symmetrical Optimum method for PI and PID controllers, Automatica, 1999, Vol. 35, pp. 1731-1736

[20] Hjalmarsson, H., S. Gunnarsson, and M. Gevers: A convergent iterative restricted complexity control design scheme, Proceedings of the 33rd IEEE Conference on Decision and Control, Lake Buena Vista, FL, 1994, pp. 1735-1440

[21] Hjalmarsson, H., M. Gevers, S. Gunnarsson, and O. Lequin: Iterative Feedback Tuning: theory and applications, IEEE Control Systems Magazine, 1998, Vol. 18, pp. 26-41

[22] Lequin, O., M. Gevers, M. Mossberg, E. Bosmans, and L. Triest: Iterative Feedback Tuning of PID parameters: comparison with classical tuning rules, Control Engineering Practice, 2003, Vol. 11, pp. 1023-1033

[23] Hamamoto, K., T. Fukuda, and T. Sugie: Iterative Feedback Tuning of controllers for a two-mass-spring system with friction, Control Engineering Practice, 2003, Vol. 11, pp. 1061-1068

[24] Rosenwasser, E. and R. Yusupov: Sensitivity of Automatic Control Systems, CRC Press, Boca Raton, FL, 2000

[25] Kovács, Sz. and L. T. Kóczy: Application of an approximate fuzzy logic controller inan AGV steering system, path tracking and collision avoidance strategy, Fuzzy Set Theory, Tatra Mountains Mathematical Publications, 1999, Vol. 16, pp. 456-467

[26] Tar, J. K., I. J. Rudas, J. F. Bitó, L. Horváth, and K. Kozlowski: Analysis of the effect ot backlash and joint acceleration measurement noise in the adaptive control of electro-mechanical systems, Proceedings of the ISIE 2003 International Symposium on Industrial Electronics, Rio de Janeiro, Brasil, 2003, CD issue, file BF-000965.pdf

[27] Vaščák. J. and L. Madarász: Automatic adaptation of fuzzy controllers, Acta Polytechnica Hungarica, 2005, Vol. 2, No. 2, pp. 5-18

Frequent Pattern Mining in Web Log Data

Renáta Iváncsy, István Vajk

Department of Automation and Applied Informatics,
and HAS-BUTE Control Research Group
Budapest University of Technology and Economics
Goldmann Gy. tér 3, H-1111 Budapest, Hungary
e-mail: {renata.ivancsy, vajk}@aut.bme.hu

Abstract: Frequent pattern mining is a heavily researched area in the field of data mining with wide range of applications. One of them is to use frequent pattern discovery methods in Web log data. Discovering hidden information from Web log data is called Web usage mining. The aim of discovering frequent patterns in Web log data is to obtain information about the navigational behavior of the users. This can be used for advertising purposes, for creating dynamic user profiles etc. In this paper three pattern mining approaches are investigated from the Web usage mining point of view. The different patterns in Web log mining are page sets, page sequences and page graphs.

Keywords: Pattern mining, Sequence mining, Graph Mining, Web log mining

1 Introduction

The expansion of the World Wide Web (Web for short) has resulted in a large amount of data that is now in general freely available for user access. The different types of data have to be managed and organized in such a way that they can be accessed by different users efficiently. Therefore, the application of data mining techniques on the Web is now the focus of an increasing number of researchers. Several data mining methods are used to discover the hidden information in the Web. However, Web mining does not only mean applying data mining techniques to the data stored in the Web. The algorithms have to be modified such that they better suit the demands of the Web. New approaches should be used which better fit the properties of Web data. Furthermore, not only data mining algorithms, but also artificial intelligence, information retrieval and natural language processing techniques can be used efficiently. Thus, Web mining has been developed into an autonomous research area.

The focus of this paper is to provide an overview how to use frequent pattern mining techniques for discovering different types of patterns in a Web log database. The three patterns to be searched are frequent itemsets, sequences and tree patterns. For each of the problem an algorithm was developed in order to discover the patterns efficiently. The frequent itemsets (frequent page sets) are discovered using the ItemsetCode algorithm presented in [1]. The main advantage of the ItemsetCode algorithm is that it discovers the small frequent itemsets in a very quick way, thus the task of discovering the longer ones is enhanced as well. The algorithm that discovers the frequent page sequences is called SM-Tree algorithm [2] and the algorithm that discovers the tree-like patters is called PD-Tree algorithm [3]. Both of the algorithms exploit the benefit of using automata theory approach for discovering the frequent patterns. The SM-Tree algorithm uses state machines for discovering the sequences, and the PD-Tree algorithm uses pushdown automatons for determining the support of the tree patterns in a tree database.

The organization of the paper is as follows. Section 2 introduces the basic tasks of Web mining. In Section 3 the Web usage mining is described in detail. The different tasks in the process of Web usage mining is depicted as well. Related Work can be found in Section 4. The algorithms used in the pattern discovery phase of the mining process are described briefly in Section 5. The preprocessing steps are described in Section 6. The results of the mining process can be found in Section 7.

2 Web Mining Approaches

Web mining involves a wide range of applications that aims at discovering and extracting hidden information in data stored on the Web. Another important purpose of Web mining is to provide a mechanism to make the data access more efficiently and adequately. The third interesting approach is to discover the information which can be derived from the activities of users, which are stored in log files for example for predictive Web caching [4]. Thus, Web mining can be categorized into three different classes based on which part of the Web is to be mined [5,6,7]. These three categories are (i) Web content mining, (ii) Web structure mining and (iii) Web usage mining. For detailed surveys of Web mining please refer to [5,6,8,9].

Web content mining [10,9] is the task of discovering useful information available on-line. There are different kinds of Web content which can provide useful information to users, for example multimedia data, structured (i.e. XML documents), semi-structured (i.e. HTML documents) and unstructured data (i.e. plain text). The aim of Web content mining is to provide an efficient mechanism to help the users to find the information they seek. Web content mining includes the task of organizing and clustering the documents and providing search engines for accessing the different documents by keywords, categories, contents etc.

Web structure mining [11,12,13,14] is the process of discovering the structure of hyperlinks within the Web. Practically, while Web content mining focuses on the inner-document information, Web structure mining discovers the link structures at the inter-document level. The aim is to identify the authoritative and the hub pages for a given subject. Authoritative pages contain useful information, and are supported by several links pointing to it, which means that these pages are highly-referenced. A page having a lot of referencing hyperlinks means that the content of the page is useful, preferable and maybe reliable. Hubs are Web pages containing many links to authoritative pages, thus they help in clustering the authorities. Web structure mining can be achieved only in a single portal or also on the whole Web. Mining the structure of the Web supports the task of Web content mining. Using the information about the structure of the Web, the document retrieval can be made more efficiently, and the reliability and relevance of the found documents can be greater. The graph structure of the web can be exploited by Web structure mining in order to improve the performance of the information retrieval and to improve classification of the documents.

Web usage mining is the task of discovering the activities of the users while they are browsing and navigating through the Web. The aim of understanding the navigation preferences of the visitors is to enhance the quality of electronic commerce services (e-commerce), to personalize the Web portals [15] or to improve the Web structure and Web server performance [16]. In this case, the mined data are the log files which can be seen as the secondary data on the web where the documents accessible through the Web are understood as primary data.

There are three types of log files that can be used for Web usage mining. Log files are stored on the server side, on the client side and on the proxy servers. By having more than one place for storing the information of navigation patterns of the users makes the mining process more difficult. Really reliable results could be obtained only if one has data from all three types of log file. The reason for this is that the server side does not contain records of those Web page accesses that are cached on the proxy servers or on the client side. Besides the log file on the server, that on the proxy server provides additional information. However, the page requests stored in the client side are missing. Yet, it is problematic to collect all the information from the client side. Thus, most of the algorithms work based only the server side data. Some commonly used data mining algorithms for Web usage mining are association rule mining, sequence mining and clustering [17].

3 Web Usage Mining

Web usage mining, from the data mining aspect, is the task of applying data mining techniques to discover usage patterns from Web data in order to understand and better serve the needs of users navigating on the Web [18]. As every data mining task, the process of Web usage mining also consists of three main steps: (i) preprocessing, (ii) pattern discovery and (iii) pattern analysis.

In this work pattern discovery means applying the introduced frequent pattern discovery methods to the log data. For this reason the data have to be converted in the preprocessing phase such that the output of the conversion can be used as the input of the algorithms. Pattern analysis means understanding the results obtained by the algorithms and drawing conclusions.

Figure 1 shows the process of Web usage mining realized as a case study in this work. As can be seen, the input of the process is the log data. The data has to be preprocessed in order to have the appropriate input for the mining algorithms. The different methods need different input formats, thus the preprocessing phase can provide three types of output data.

The frequent patterns discovery phase needs only the Web pages visited by a given user. In this case the sequences of the pages are irrelevant. Also the duplicates of the same pages are omitted, and the pages are ordered in a predefined order.

In the case of sequence mining, however, the original ordering of the pages is also important, and if a page was visited more than once by a given user in a user-defined time interval, then it is relevant as well. For this reason the preprocessing module of the whole system provides the sequences of Web pages by users or user sessions.

For subtree mining not only the sequences are needed but also the structure of the web pages visited by a given user. In this case the backward navigations are omitted; only the forward navigations are relevant, which form a tree for each user. After the discovery has been achieved, the analysis of the patterns follows. The whole mining process is an iterative task which is depicted by the feedback in Figure 1. Depending on the results of the analysis either the parameters of the preprocessing step can be tuned (i.e. by choosing another time interval to determine the sessions of the users) or only the parameters of the mining algorithms. (In this case that means the minimum support threshold.)

In the case study presented in this work the aim of Web usage mining is to discover the frequent pages visited at the same time, and to discover the page sequences visited by users. The results obtained by the application can be used to form the structure of a portal satisfactorily for advertising reasons and to provide a more personalized Web portal.

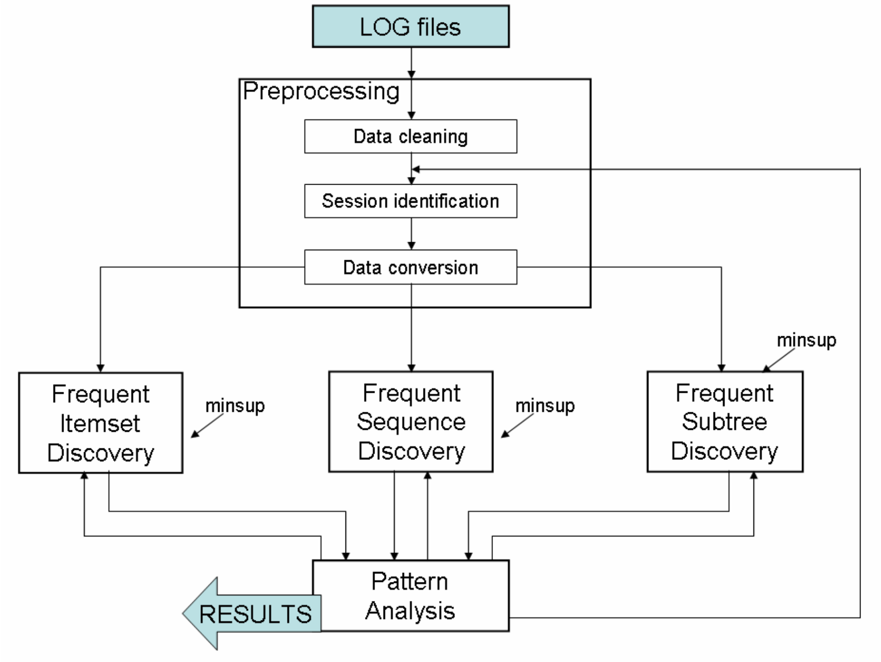


Figure 1
Process of Web usage mining

4 Related Work

In Web usage mining several data mining techniques can be used. Association rules are used in order to discover the pages which are visited together even if they are not directly connected, which can reveal associations between group of users with specific interest [15]. This information can be used for example for restructuring Web sites by adding links between those pages which are visited together. Association rules in Web logs are discovered in [19,20,21,22,23]. Sequence mining can be used for discover the Web pages which are accessed immediately after another. Using this knowledge the trends of the activity of the users can be determined and predictions to the next visited pages can be calculated. Sequence mining is accomplished in [16], where a so-called WAP-tree is used for storing the patterns efficiently. Tree-like topology patterns and frequent path traversals are searched by [19,24,25,26].

Web usage mining is elaborated in many aspects. Besides applying data mining techniques also other approaches are used for discovering information. For example [7] uses probabilistic grammar-based approach, namely an Ngram model for capturing the user navigation behavior patterns. The Ngram model assumes that the last *N* pages browsed affect the probability of identifying the next page to be visited. [27] uses Probabilistic Latent Semantic Analysis (PLSA) to discover the navigation patterns. Using PLSA the hidden semantic relationships among users and between users and Web pages can be detected. In [28] Markov assumptions are used as the basis to mine the structure of browsing patterns. For Web prefetching [29] uses Web log mining techniques and [30] uses a Markov predictor.

5 Overview of the Mining Algorithms

Before investigating the whole process of Web usage mining, and before explaining the important steps of the process, the frequent pattern mining algroithms are explained hier briefly. It is necessary to understant the mechanism of these algorithms in order to understand their results. Another important aspect is to determine the input parameters of the algorithm in order to hav the opportunity of providing the adequate input formats by the preprocessing phase of the mining process.

As mentioned earlier the frequent set of pages are discovered using the ItemsetCode algorithm [1]. It is a level-wise “candidate generate and test” method that is based only partionally on the Apriori hypothesis. The aim of the algorithm is to enhance the Apriori algorithm on the low level. It means, enhancing the step of discoverint the small frequent itemsets. In such a way also the greater itemsets are discovered more quickly. The idea of the ItemsetCode algorithm is to is to reduce the problem of discovering the 3 and 4-frequent itemsets back to the problem of discovering 2-frequent itemsets by using a coding mechanism. The ItemsetCode algorithm discovers the 1 and 2-frequent itemsets in the quickest way by directly indexing a matrix. The 2-frequent itemsets are coded and the 3 and 4-candidates are created by pairing the codes. The counters for the 3 and 4-candidates are stored in a jugged array in order to have a storage structure of moderate memory requirements. The way in which the candidates are created enables us to use the jugged array in a very efficient way by using two indirections only. Furthermore, the memory requirement of the structure is also low. The algorithm only partially exploits the benefits of the Apriori hypothesis. The reason is the compact storage structure for the candidates. The ItemsetCode algorithm discovers the large itemset efficiently because of the quick discovery of the small itemsets. Its level-wise approach ensures the fact that its memory requirement does not depend on the number of transactions. The input format of the ItemsetCode algorithm suits the input format of other frequent mining algorithms. It reads the transactions by rows and each row contains the list of items.

The page sequences are discovered using the SM-Tree algotihm (State Machine-Tree algorithm) [2]. The main idea of the SM-Tree algorithm is to test the subsequence inclusion in such a way that the items of the input sequence are processed exactly once. The basis of the new approach is the deterministic finite state machines created for the candidates. By joining the several automatons a new structure called SM-Tree is created such that handling a large number of candidates is faster than in the case of using different state machines for each candidate. Based on its features the SM-Tree structure can be handled efficiently. This can be done by exploiting the benefits of having two types of states, namely the fixed and the temporary states. The further benefit of the suggested algorithm is that its memory requirement is independent from the number of transactions which comes from the level-wise approach. The input format of the SM-Tree algorithms contains rows of transactions, where each row contains a sequence, where the itemsets are separeted by a -1 value.

The PD-Tree algorithm proposes a new method for determining whether a tree is contained by another tree. This can be done by using a pushdown automaton. In order to provide an input to the automaton, the tree is represented as a string. For handling the large number of candidates eficiently the join operation between the automatons were proposed, and the resulting new structure is called PD-Tree. The new structure makes it possible to discover the support of each candidate at the same time by processing the items of a transaction exactly once. The benefit of the PD-Tree is that it uses only one stack to accomplish the mining process. Experimental results show the time saving when using the PD-Tree instead of using several pushdown automatons. The input format of the algorithm also contains rows of transactions where each transaction contain a tree. A tree is represented with strings.

6 Data Preprocessing

The data in the log files of the server about the actions of the users can not be used for mining purposes in the form as it is stored. For this reason a preprocessing step must be performed before the pattern discovering phase.

The preprocessing step contains three separate phases. Firstly, the collected data must be cleaned, which means that graphic and multimedia entries are removed. Secondly, the different sessions belonging to different users should be identified. A session is understood as a group of activities performed by a user when he is navigating through a given site. To identify the sessions from the raw data is a complex step, because the server logs do not always contain all the information needed. There are Web server logs that do not contain enough information to reconstruct the user sessions, in this case for example time-oriented heuristics can be used as described in [31]. After identifying the sessions, the Web page sequences are generated which task belongs to the first step of the preprocessing. The third step is to convert the data into the format needed by the mining algorithms. If the sessions and the sequences are identified, this step can be accomplished more easily.

In our experiments we used two web server log files, the first one was the *msnbc.com anonymous data*[1] and the second one was a Click Stream data downloaded from the *ECML/PKDD 2005 Discovery Challenge*[2]. Both of the log files are in different formats, thus different preprocessing steps were needed.

The msnbc log data describes the page visits of users who visited msnbc.com on September 28, 1999. Visits are recorded at the level of URL category and are recorded in time order. This means that in this case the first phase of the preprocessing step can be omitted. The data comes from Internet Information Server (IIS) logs for msnbc.com. Each row in the dataset corresponds to the page visits of a user within a twenty-four hour period. Each item of a row corresponds to a request of a user for a page. The pages are coded as shown in Table 1. The client-side cached data is not recorded, thus this data contains only the server-side log.

Table 1

Codes for the msnbc.com page categories

category	code	category	code	category	code
frontpage	1	misc	7	summary	13
news	2	weather	8	bbs	14
tech	3	health	9	travel	15
local	4	living	10	msn-news	16
opinion	5	business	11	msn-sport	17
On-air	6	sports	12		

In the case of the msnbc data only the rows have to be converted into itemsets, sequences and trees. The other preprocessing steps are done already. A row is converted into an itemset by omitting the duplicates of the pages, and sorting them regarding their codes. In this way the ItemsetCode algorithm can be executed easily on the dataset.

In order to have sequence patterns the row has to be converted such that they represent sequences. A row corresponds practically to a sequence having only one item in each itemset. Thus converting a row into the sequence format needed by the SM-Tree algorithm means to insert a -1 between each code.

[1] http://kdd.ics.uci.edu/databases/msnbc/msnbc.html

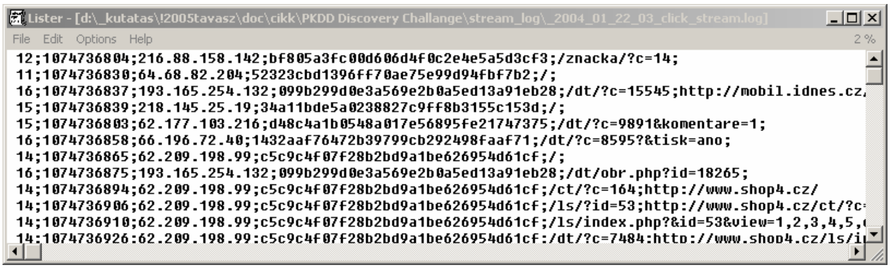
[2] http://lisp.vse.cz/challenge/CURRENT/

In order to have the opportunity mining tree-like patterns the database has to be converted such that the transactions represent trees. For this reason each row is processed in the following way. The root of the tree is the first item of the row. From the subsequent items a branch is created until an item is reached which was already inserted into the tree. In this case the algorithm inserts as many -1 item into the string representation of the tree as the number of the items is between the new item and the previous occurrence of the same item. The further items form another branch in the tree. For example given the row: “1 2 3 4 2 5” then the tree representation of the row is the following: “1 2 3 4 -1 -1 5”.

In case of the Click Stream data, the preprocessing phase needs more work. It contains 546 files where each file contains the information collected during one hour from the activities of the users in a Web store. Each row of the log contains the following parts:

- a shop identifier
- time
- IP address
- automatic created unique session identifier
- visited page
- referrer

In Figure 2 a part of the raw log file can be observed. Because in this case the sessions have already been identified in the log file, the Web page sequences for the same sessions have to be collected only in the preprocessing step. This can be done in the different files separately, or through all the log files. After the sequences are discovered, the different web pages are coded, and similarly to the msnbc data, the log file has to be converted into itemsets and sequences.


Lister - [d:_kutatas\!2005tavasz\doc\cikk\PKDD Discovery Challange\stream_log_2004_01_22_03_click_stream.log]
File
Edit
Options
Help
2 %

```
12;1074736804;216.88.158.142;bf805a3fc00d606d4f0c2e4e5a5d3cf3;/znacka/?c=14;
11;1074736830;64.68.82.204;52323cbd1396ff70ae75e99d94fbf7b2;/;
16;1074736837;193.165.254.132;099b299d0e3a569e2b0a5ed13a91eb28;/dt/?c=15545;http://mobil.idnes.cz/
15;1074736839;218.145.25.19;34a11bde5a0238827c9ff8b3155c153d;/;
15;1074736803;62.177.103.216;d48c4a1b0548a017e56895fe21747375;/dt/?c=9891&komentare=1;
16;1074736858;66.196.72.40;1432aaf76472b39799cb292498faaf71;/dt/?c=8595?&tisk=ano;
14;1074736865;62.209.198.99;c5c9c4f07f28b2bd9a1be626954d61cf;/;
16;1074736875;193.165.254.132;099b299d0e3a569e2b0a5ed13a91eb28;/dt/obr.php?id=18265;
14;1074736894;62.209.198.99;c5c9c4f07f28b2bd9a1be626954d61cf;/ct/?c=164;http://www.shop4.cz/
14;1074736906;62.209.198.99;c5c9c4f07f28b2bd9a1be626954d61cf;/ls/?id=53;http://www.shop4.cz/ct/?c=
14;1074736910;62.209.198.99;c5c9c4f07f28b2bd9a1be626954d61cf;/ls/index.php?&id=53&view=1,2,3,4,5,
14;1074736926;62.209.198.99;c5c9c4f07f28b2bd9a1be626954d61cf;/dt/?c=7484;http://www.shop4.cz/ls/i
```

Figure 2

An example of raw log file

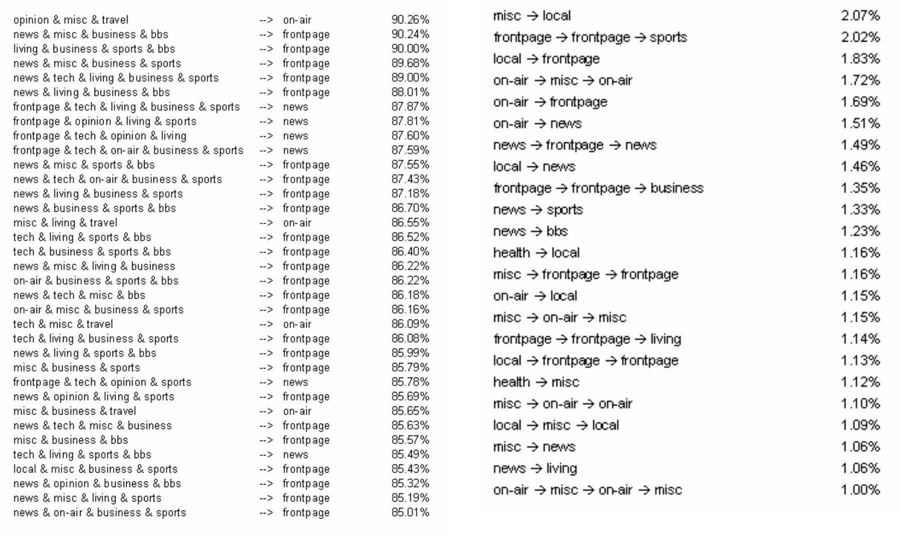
7 Data Mining and Pattern Analysis

As it is depicted in Figure 1, the Web usage mining system is able to use all three frequent pattern discovery tasks described in this work. For the mining process, besides the input data, the minimum support threshold value is needed. It is one of the key issues, to which value the support threshold should be set. The right answer can be given only with the user interactions and many iterations until the appropriate values have been found. For this reason, namely, that the interaction of the users is needed in this phase of the mining process, it is advisable executing the frequent pattern discovery algorithm iteratively on a relatively small part of the whole dataset only. Choosing the right size of the sample data, the response time of the application remains small, while the sample data represents the whole data accurately. Setting the minimum support threshold parameter is not a trivial task, and it requires a lot of practice and attention on the part of the user.

The frequent itemset discovery and the association rule mining was accomplished using the ItemsetCode algorithm with different minimum support and minimum confidence threshold values. Figure 3 (a) depicts the association rules generated from msnbc.com data at a minimum support threshold of 0.1% and at a minimum confidence threshold of 85% (which is depicted in the figure). Analyzing the results, one can make the advertising process more successful and the structure of the portal can be changed such that the pages contained by the rules are accessible from each other.

Another type of decision can be made based on the information gained from a sequence mining algorithm. Figure 3 (b) shows a part of the discovered sequences of the SM-Tree algorithm from the msnbc.com data. The percentage values depicted in Figure 3 (b) are the support of the sequences.

The frequent tree mining task was accomplished using the PD-Tree algorithm. A part of the result of the tree mining algorithm is depicted in Figure 4 (a). The patterns contain beside the trees (represented in string format), also the support values. The graphical representations of the patterns are depicted in Figure 4 (b) without the support values.



opinion & misc & travel	-->	on-air	90.26%
news & misc & business & bbs	-->	frontpage	90.24%
living & business & sports & bbs	-->	frontpage	90.00%
news & misc & business & sports	-->	frontpage	89.68%
news & tech & living & business & sports	-->	frontpage	89.00%
news & living & business & bbs	-->	frontpage	88.01%
frontpage & tech & living & business & sports	-->	news	87.87%
frontpage & opinion & living & sports	-->	news	87.81%
frontpage & tech & opinion & living	-->	news	87.60%
frontpage & tech & on-air & business & sports	-->	news	87.59%
news & misc & sports & bbs	-->	frontpage	87.55%
news & tech & on-air & business & sports	-->	frontpage	87.43%
news & living & business & sports	-->	frontpage	87.18%
news & business & sports & bbs	-->	frontpage	86.70%
misc & living & travel	-->	on-air	86.55%
tech & living & sports & bbs	-->	frontpage	86.52%
tech & business & sports & bbs	-->	frontpage	86.40%
news & misc & living & business	-->	frontpage	86.22%
on-air & business & sports & bbs	-->	frontpage	86.22%
news & tech & misc & bbs	-->	frontpage	86.18%
on-air & misc & business & sports	-->	frontpage	86.16%
tech & misc & travel	-->	on-air	86.09%
tech & living & business & sports	-->	frontpage	86.08%
news & living & sports & bbs	-->	frontpage	85.99%
misc & business & sports	-->	frontpage	85.79%
frontpage & tech & opinion & sports	-->	news	85.78%
news & opinion & living & sports	-->	frontpage	85.69%
misc & business & travel	-->	on-air	85.65%
news & tech & misc & business	-->	frontpage	85.63%
misc & business & bbs	-->	frontpage	85.57%
tech & living & sports & bbs	-->	news	85.49%
local & misc & business & sports	-->	frontpage	85.43%
news & opinion & business & bbs	-->	frontpage	85.32%
news & misc & living & sports	-->	frontpage	85.19%
news & on-air & business & sports	-->	frontpage	85.01%

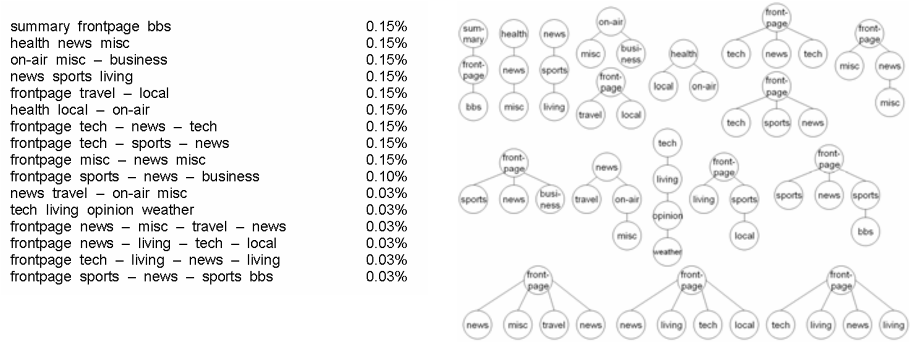
(a)

misc → local	2.07%
frontpage → frontpage → sports	2.02%
local → frontpage	1.83%
on-air → misc → on-air	1.72%
on-air → frontpage	1.69%
on-air → news	1.51%
news → frontpage → news	1.49%
local → news	1.46%
frontpage → frontpage → business	1.35%
news → sports	1.33%
news → bbs	1.23%
health → local	1.16%
misc → frontpage → frontpage	1.16%
on-air → local	1.15%
misc → on-air → misc	1.15%
frontpage → frontpage → living	1.14%
local → frontpage → frontpage	1.13%
health → misc	1.12%
misc → on-air → on-air	1.10%
local → misc → local	1.09%
misc → news	1.06%
news → living	1.06%
on-air → misc → on-air → misc	1.00%

(b)

Figure 3

(a) Association rules and (b) sequential rules based on the msnbc.data



summary frontpage bbs	0.15%
health news misc	0.15%
on-air misc – business	0.15%
news sports living	0.15%
frontpage travel – local	0.15%
health local – on-air	0.15%
frontpage tech – news – tech	0.15%
frontpage tech – sports – news	0.15%
frontpage misc – news misc	0.15%
frontpage sports – news – business	0.10%
news travel – on-air misc	0.03%
tech living opinion weather	0.03%
frontpage news – misc – travel – news	0.03%
frontpage news – living – tech – local	0.03%
frontpage tech – living – news – living	0.03%
frontpage sports – news – sports bbs	0.03%

(a) (b)

Figure 4

Frequent tree patterns based on msnbc.data in (a) string and (b) graphical representation

Conclusions

This paper deals with the problem of discovering hidden information from large amount of Web log data collected by web servers. The contribution of the paper is to introduce the process of web log mining, and to show how frequent pattern discovery tasks can be applied on the web log data in order to obtain useful information about the user's navigation behavior.

Acknowledgement

This work has been supported by the Mobile Innovation Center, Hungary, by the fund of the Hungarian Academy of Sciences for control research and the Hungarian National Research Fund (grant number: T042741).

References

[1] R. Iváncsy and I. Vajk, "Time- and Memory-Efficient Frequent Itemset Discovering Algorithm for Association Rule Mining." *International Journal of Computer Applications in Techology, Special Issue on Data Mining Applications (in press)*

[2] R. Iváncsy and I. Vajk, "Efficient Sequential Pattern Mining Algorithms." WSEAS Transactions on Computers, Vol. 4, Num. 2, 2005, pp. 96-101

[3] R. Iváncsy and I. Vajk, "PD-Tree: A New Approach to Subtree Discovery.", WSEAS Transactions on Information Science and Applications, Vol. 2, Num. 11, 2005, pp. 1772-1779

[4] Q. Yang and H. H. Zhang, "Web-log mining for predictive web caching." *IEEE Trans. Knowl. Data Eng.*, Vol. 15, No. 4, pp. 1050-1053, 2003

[5] Kosala and Blockeel, "Web mining research: A survey," *SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery and Data Mining, ACM*, Vol. 2, 2000

[6] S. K. Madria, S. S. Bhowmick, W. K. Ng, and E.-P. Lim, "Research issues in web data mining," in *Data Warehousing and Knowledge Discovery*, 1999, pp. 303-312

[7] J. Borges and M. Levene, "Data mining of user navigation patterns," in WEBKDD, 1999, pp. 92-111

[8] M. N. Garofalakis, R. Rastogi, S. Seshadri, and K. Shim, "Data mining and the web: Past, present and future," in *ACM CIKM'99 2nd Workshop on Web* Information *and Data Management (WIDM'99), Kansas City, Missouri, USA, November 5-6, 1999*, C. Shahabi, Ed. ACM, 1999, pp. 43-47

[9] S. Chakrabarti, "Data mining for hypertext: A tutorial survey." *SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on* Knowledge *Discovery and Data Mining, ACM*, Vol. 1, No. 2, pp. 1-11, 2000

[10] M. Balabanovic and Y. Shoham, "Learning information retrieval agents: Experiments with automated web browsing," in *Proceedings of the AAAI Spring Symposium on Information Gathering from Heterogenous, Distributed Resources*, 1995, pp. 13-18

[11] S. Chakrabarti, B. Dom, and P. Indyk, "Enhanced hypertext categorization using hyperlinks," in *SIGMOD '98: Proceedings of the 1998 ACM*

SIGMOD international conference on Management of data. New York, NY, USA: ACM Press, 1998, pp. 307-318

[12] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins, "The Web as a graph: Measurements, models and methods," *Lecture Notes in Computer Science*, Vol. 1627, pp. 1-18, 1999

[13] J. Hou and Y. Zhang, "Effectively finding relevant web pages from linkage information." *IEEE Trans. Knowl. Data Eng.*, Vol. 15, No. 4, pp. 940-951, 2003

[14] H. Han and R. Elmasri, "Learning rules for conceptual structure on the web," *J. Intell. Inf. Syst.*, Vol. 22, No. 3, pp. 237-256, 2004

[15] M. Eirinaki and M. Vazirgiannis, "Web mining for web personalization," *ACM* Trans. *Inter. Tech.*, Vol. 3, No. 1, pp. 1-27, 2003

[16] J. Pei, J. Han, B. Mortazavi-Asl, and H. Zhu, "Mining access patterns efficiently from web logs," in *PADKK '00: Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications*. London, UK: Springer-Verlag, 2000, pp. 396-407

[17] R. Cooley, B. Mobasher, and J. Srivastava, "Data preparation for mining world wide web browsing patterns," *Knowledge and Information Systems*, Vol. 1, No. 1, pp. 5-32, 1999

[18] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, "Web usage mining: Discovery and applications of usage patterns from web data," *SIGKDD Explorations*, Vol. 1, No. 2, pp. 12-23, 2000

[19] M. S. Chen, J. S. Park, and P. S. Yu, "Data mining for path traversal patterns in a web environment," in *Sixteenth International Conference on* Distributed *Computing Systems*, 1996, pp. 385-392

[20] J. Punin, M. Krishnamoorthy, and M. Zaki, "Web usage mining: Languages and algorithms," in *Studies in Classification, Data Analysis, and* Knowledge *Organization*. Springer-Verlag, 2001

[21] P. Batista, M. ario, and J. Silva, "Mining web access logs of an on-line newspaper," 2002

[22] O. R. Zaiane, M. Xin, and J. Han, "Discovering web access patterns and trends by applying olap and data mining technology on web logs," in *ADL '98: Proceedings of the Advances in Digital Libraries Conference*. Washington, DC, USA: IEEE Computer Society, 1998, pp. 1-19

[23] J. F. F. M. V. M. Li Shen, Ling Cheng and T. Steinberg, "Mining the most interesting web access associations," in *WebNet 2000-World Conference on the WWW and Internet*, 2000, pp. 489-494

[24] X. Lin, C. Liu, Y. Zhang, and X. Zhou, “Efficiently computing frequent tree-like topology patterns in a web environment,” in TOOLS ’99: Proceedings of the 31st International Conference on Technology of Object-Oriented Language and Systems. Washington, DC, USA: IEEE Computer Society, 1999, p. 440

[25] A. Nanopoulos and Y. Manolopoulos, “Finding generalized path patterns for web log data mining,” in ADBIS-DASFAA ’00: Proceedings of the East-European Conference on Advances in Databases and Information Systems Held Jointly with International Conference on Database Systems for Advanced Applications. London, UK: Springer-Verlag, 2000, pp. 215-228

[26] A. Nanopoulos and Y. Manolopoulos, “Mining patterns from graph traversals,” Data and Knowledge Engineering, Vol. 37, No. 3, pp. 243-266, 2001

[27] X. Jin, Y. Zhou, and B. Mobasher, “Web usage mining based on probabilistic latent semantic analysis,” in KDD ’04: Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: ACM Press, 2004, pp. 197-205

[28] S. Jespersen, T. B. Pedersen, and J. Thorhauge, “Evaluating the markov assumption for web usage mining,” in WIDM ’03: Proceedings of the 5th ACM international workshop on Web information and data management. New York, NY, USA: ACM Press, 2003, pp. 82-89

[29] A. Nanopoulos, D. Katsaros, and Y. Manolopoulos, “Exploiting web log mining for web cache enhancement,” in WEBKDD ’01: Revised Papers from the Third International Workshop on Mining Web Log Data Across All Customers Touch Points. London, UK: Springer-Verlag, 2002, pp. 68-87

[30] A. Nanopoulos, D. Katsaros and Y. Manolopoulos, “A data mining algorithm for generalized web prefetching,” IEEE Transactions on Knowledge and Data Engineering, Vol. 15, No. 5, pp. 1155-1169, 2003

[31] J. Zhang and A. A. Ghorbani, “The reconstruction of user sessions from a server log using improved timeoriented heuristics.” in CNSR. IEEE Computer Society, 2004, pp. 315-322

A Brief Survey and Comparison on Various Interpolation Based Fuzzy Reasoning Methods

Zsolt Csaba Johanyák [1], Szilveszter Kovács [2]

[1] Department of Information Technology, GAMF Faculty, Kecskemét College, H-6001 Kecskemét, Pf. 91, johanyak.csaba@gamf.kefo.hu

[2] Department of Information Technology, University of Miskolc, Miskolc-Egyetemváros, H-3515 Miskolc, Hungary, szkovacs@iit.uni-miskolc.hu

Abstract: Fuzzy systems based on sparse rule bases produce the conclusion through approximation. This paper is the first part of a longer survey that aims to provide a qualitative view through the presentation of the basic ideas and characteristics of some methods and defining a general condition set brought together from an application-oriented point of view.

Keywords: interpolative fuzzy reasoning, sparse rule base

1 Why do We Need Interpolation?

The functioning of systems working with fuzzy logic is based on rules. The rule base is considered dense when for all the possible observations there exists at least one rule, whose antecedent part overlaps the input data, at least partially. Otherwise, the rule base is considered to be sparse. The classical inference methods (e.g. compositional rule of inference) are not able to produce an output for the observations covered by none of the rules. That is why the systems based on a sparse rule base should adopt inference techniques, which in the lack of matching rules perform an approximate reasoning taking into consideration the existing rules. The most often used methods for this purpose are called interpolative methods.

2 General Conditions on Rule Interpolation Methods

A unified condition system related to the interpolative methods would make the evaluation and comparison of the different techniques based on the same fundamentals possible. However, according to the existing literature (e.g. [1] [6] [16] [18]) can be found only partly consistent conditions and condition groups, which are put together taking different points of view into consideration. Therefore, as a step towards the unification, the conditions considered to be the most relevant ones from the application-oriented aspects are going to be reviewed and based on them, some of the well known methods are going to be compared in the followings.

General conditions on rule interpolation methods:

1 *Avoidance of the abnormal conclusion* [1] [6] [16]. The estimated fuzzy set should be a valid one. This requisite can be described by the constraints (1) and (2) according to [16].

$$\inf\left\{B^*_\alpha\right\} \leq \sup\left\{B^*_\alpha\right\} \qquad \forall \alpha \in [0,1] \tag{1}$$

$$\inf\left\{B^*_{\alpha 1}\right\} \leq \inf\left\{B^*_{\alpha 2}\right\} \leq \sup\left\{B^*_{\alpha 2}\right\} \leq \sup\left\{B^*_{\alpha 1}\right\} \qquad \forall \alpha_1 < \alpha_2 \in [0,1] \tag{2}$$

where $\inf\left\{B^*_\alpha\right\}$ and $\sup\left\{B^*_\alpha\right\}$ are the lower and upper endpoints of the actual α-cut of the estimated fuzzy set.

2 *The continuity of the mapping between the antecedent and consequent fuzzy sets* [1] [6]. This condition indicates that similar observations should lead to similar results.

3 *Preserving the "in between"* [6]. If the antecedent sets of two neighbouring rules surround an observation, the approximated conclusion should be surrounded by the consequent sets of those rules, too.

4 *Compatibility with the rule base* [1] [6]. This means the requisite on the validity of the modus ponent, namely if an observation coincides with the antecedent part of a rule, the conclusion produced by the method should correspond to the consequent part of that rule.

5 *The fuzziness of the approximated result.* There are two opposite approaches in the literature related to this topic [18]. According to the first subcondition (5a), the less uncertain the observation is the less fuzziness should have the approximated consequent [1] [6]. With other words in case of a singleton type observation the method should produce a crisp valued consequence. The second approach (5b) originates the fuzziness of the estimated consequent from the nature of the fuzzy rule base [16]. Thus, crisp conclusion can be expected only if all the consequents of the rules taken into consideration during the interpolation are singleton shaped, i.e. the knowledge base produces certain information from fuzzy input data.

6 *Approximation capability* (*stability* [e.g. 17]). The estimated rule should approximate with the possible highest degree the relation between the

antecedent and consequent universes. If the number of the measurement (knot) points tends to infinite, the result should converge to the approximated function independently from the position of the knot points.

7 *Conserving the piece-wise linearity* [1]. If the fuzzy sets of the rules taken into consideration are piece-wise linear, the approximated sets should conserve this feature.

8 *Applicability in case of multidimensional antecedent universe*.

9 *Applicability without any constraint regarding to the shape of the fuzzy sets*. This condition can be lightened practically to the case of polygons, since piece-wise linear sets are most frequently encountered in the applications.

3 Surveying Some Interpolative Methods

The techniques being reviewed can be divided into two groups relating to their conception. The members of the first group produce the approximated conclusion from the observation directly. The second group contains methods that reach the target in two steps. In the first step they interpolate a new rule whose antecedent part overlaps the observation at least partially. The estimated conclusion is determined in the second step based on the similarity between the observation and the antecedent part of the new rule.

Further on mostly the case of the one-dimensional antecedent universes are presented for the sake of easy understanding of the key ideas of the methods. As several methods need the existence of two or more rules flanking the observation, therefore it is assumed that they exist and are known. The methods are not based on the same principles, hence sometimes they approach the topic of the rule interpolation from different viewpoints.

3.1 The Linear Interpolation Introduced by Kóczy and Hirota and the Derived Methods

The first subset of the methods producing the approximated conclusion from the observation directly contains the technique introduced by Kóczy and Hirota and those ones that have been derived from it aiming its extension and improvement. First the most famous member of this group, the KH interpolation is reviewed.

3.1.1 KH Interpolation

The key idea of the method developed by Kóczy and Hirota [8] is that the approximated conclusion divides the distance between the consequent sets of the used rules in the same proportion as the observation does the distance between the antecedents of those rules (3). This is the fundamental equation of the fuzzy rule interpolation [1]. The proportions are set up separately for the lower and upper distances in the case of each α-cut.

The development of KH method was made possible by the definition of the fuzzy distance [7] and the fact that the fuzzy sets can be decomposed into α-cuts and can be composed from α-cuts (resolution and extension principle).

$$d_{\alpha}^{i}\left(A_1,A^*\right):d_{\alpha}^{i}\left(A^*,A_2\right)=d_{\alpha}^{i}\left(B_1,B^*\right):d_{\alpha}^{i}\left(B^*,B_2\right) \tag{3}$$

where A_1, A_2 are the antecedent sets of the two flanking rules, A^* is the observation, B_1, B_2 are the consequent sets of those rules, B^* is the approximated conclusion, i can be L or U depending on lower or upper type of the distance. The technique adopted for the determination of the consequent is an extension of the classic Shepard interpolation [12] for case of the fuzzy sets. The method requires the following preconditions to be fulfilled: the sets have to be convex and normal with bounded support, and at least a partial ordering should exist between the elements of the universes of discourses. The latter one is needed for the definition of the fuzzy distance.

The most important advantage of the KH interpolation is its low computational complexity that ensures the fastness required by real time applications. Its detailed analysis e.g. [10] [9] [13] led to the conclusion that the result can not be interpreted always as a fuzzy set, because e.g. by some α-cuts of the estimated consequent the lower value can be higher than the upper one (Fig. 1). The above listed publications defined application conditions that enabled the avoidance of the abnormal conclusion.

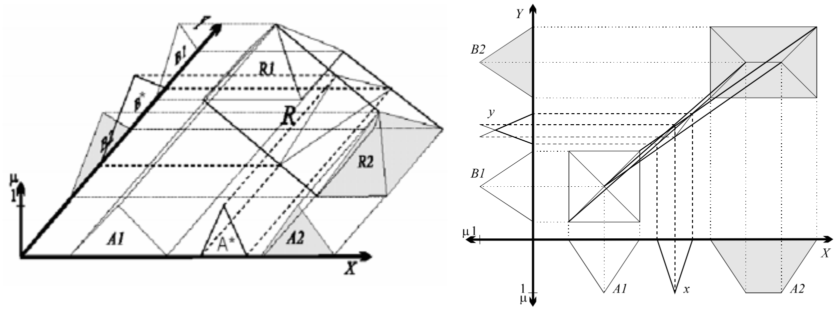


Figure 1
KH interpolation

Theoretically, an infinite number of α-cuts are needed for the exact result if there are no conditions related to the shape of the sets. However, in practice driven by need for efficiency mostly piece-wise linear generally triangle shaped or trapezoidal sets can be found, because these can be easily described by a few characteristic points. Thus supposing the method preserves the linearity completing the calculations for a finite small number of α-cuts could be enough. Although the preceding assumption is not fulfilled, in most of the applications it does not matter because of the negligible amount of the deviation [10] [9] [16].

The KH method was developed for one-dimensional antecedent universes. However, it can be applied in multi-dimensional case using distances calculated in Minkowski sense. It can be simply proven that this technique fulfils *conditions* 3, 4, 5b and 8. The stabilized (general) KH interpolation [17] also satisfies the condition 6.

The recognition of the shortcomings of the KH interpolation has led to the development of many techniques, which modified or improved the original one or offered a solution for the task of the interpolation using very new approaches. Further on some methods improving the KH technique are reviewed emphasizing those properties which are considered to be the most important.

3.1.2 Extended KH Interpolation

Several versions of the KH interpolations were developed which allow taking into consideration more than two rules during the determination of the consequence. Their common feature is that the approximation capability of the technique is getting better with the growth of the number of the rules taken into consideration.

In [8] a technique is proposed that takes into consideration the rules weighted with e.g. the reciprocal value of the square of the distance. This approach reflects that the rules situated far away from the observation are not as important as those ones in the neighbourhood of the observation.

The authors of [17] suggest to use formulas for the calculation of endpoints of α-cuts of the approximated consequence, which contain the distance on the n^{th} power, where n is number of the antecedent dimensions.

3.1.3 The VKK Method

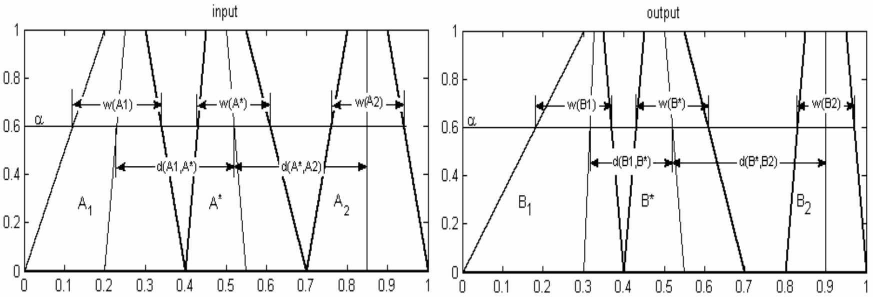


Figure 2

The method developed by Vass, Kalmár and Kóczy [19] worked out the problem of abnormal conclusion introducing modified distance measures (Fig. 2). The method is an α-cut based technique. It describes each α-cut by the position of its centre point and its width (*w*). The distance of the sets is characterized by a vector containing the Euclidean distances of the centre points (*d*). The method is also applicable in multidimensional case by calculating the resulting distance in Minkowski sense with the parameter 2 and by determining the resulting width as a geometric mean of the width values in each dimension. However, the technique cannot be applied if any of the antecedent sets taken into consideration during the interpolation are singleton shaped (none of the antecedent α-cut widths can be zero valued). Like the KH method it does not conserve the linearity, but the deviance can be proven to be negligible [1].

It can be simply proven that this technique fulfils *conditions* 3, 4, 5.a and 8.

3.1.4 Modified α-cut Based Interpolation

The modified α-cut based interpolation (MACI) [16] represents each fuzzy set by two vectors describing the left (lower) and right (upper) flanks using the technique published by Yam [21]. The vectors contain the break points in case of piece-wise linear membership functions or endpoints of predefined (usually uniform distributed) α-cuts in case of smooth membership functions. For example the antecedent set A_1 in Fig. 3 is represented by the vectors (4) and (5).

$$A_1^L = \left[x_{-1}^1, x_0^1\right] \tag{4}$$

$$A_1^R = \left[x_0^1, x_1^1\right] \tag{5}$$

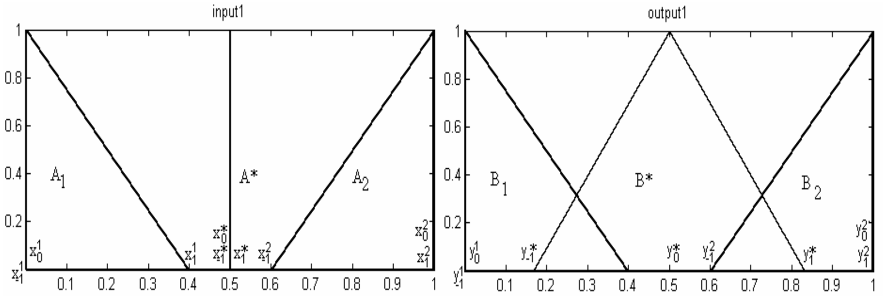


Figure 3

The graphical representation of the vectors describing the right flanks of the sets can be seen on the Figure 4. The antecedent and consequent sets are represented separately. The result will fulfil the *condition* 1 if B* is situated inside of the rectangle and above of the line l. This purpose is reached through a coordinate transformation where Z_0 is substituted by the line *l*. The approximated conclusion will be crisp only if the consequent sets of the rules taken into consideration are singletons, as well.

Although this method is not conserving the linearity, the deviance is smaller than in the case of the KH interpolation [16] and the stability experienced at the KH method [15] remains. The estimated conclusion always yields fuzziness if the consequent sets of the rules taken into consideration have fuzziness [22].

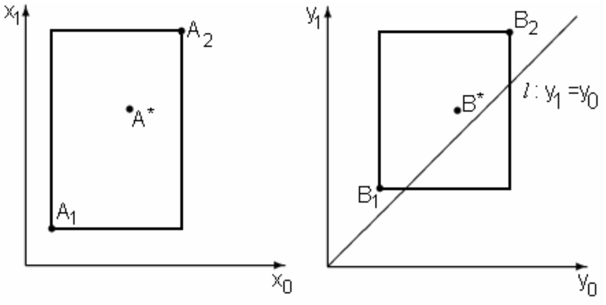


Figure 4
Graphical representation of the vectors [18]

It can be proven that the technique fulfils the *conditions* 1-4, 5b, 6, 8 and 9 with the constraint that the sets should be convex and normal. Its generalized version [14] can be used in case of non-convex fuzzy sets, too.

3.1.5 The Improved Multidimensional Modified α-cut Based Interpolation

The improved multidimensional modified α-cut based interpolation (IMUL) introduced by Wong, Gedeon and Tikk [20] combines the advantages of the MACI and the fuzziness conservation technique proposed by Kóczy and Gedeon in [3]. This method was developed for the case of multidimensional antecedent universe. The fuzzy sets are described by vectors containing the characteristic points, and the coordinate transformation introduced by MACI is used during the determination of the core of the approximated consequent.

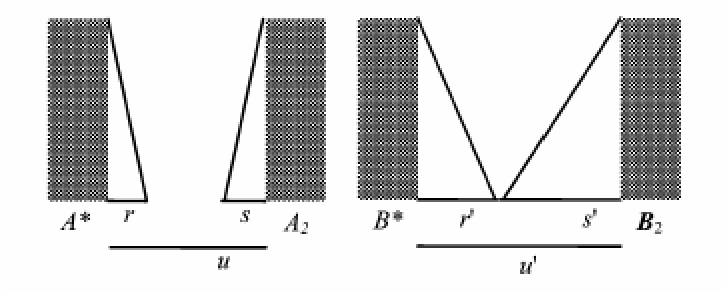


Figure 5 [23]

The fuzziness of the observation (*r*) plays a decisive role at the calculation of the flanking edges and beside this the relative fuzziness of the sets adjacent to the observation (*s/u*) and adjacent to the approximated consequent (*s'/u'*) are taken into consideration, as well. Figure 5 presents the meaning of the used notation for the case of the right flank of the conclusion.

It can be proven that the technique fulfils the *conditions* 1-4, 5a, 6, 8 and 9 with the constraint that the sets should be convex and normal.

3.2 Fuzzy Rule Interpolation in the Vague Environment

The fuzzy rule interpolation in the vague environment (FIVE) introduced by Kovács [11] puts the problem of rule interpolation in a virtual space in the so-called vague environment whose conception is based on the similarity (indistinguishability) of the objects. The similarity of two fuzzy sets in the vague environment (6) is characterized by their distance weighted with the so-called scaling function (s), which describes the vague environment. The scaling function describes the shapes of all the terms in a fuzzy partition.

$$\delta_s(x_1, x_2) = \left| \int_{x_2}^{x_1} \mathrm{s}(\mathrm{x})\mathrm{dx} \right| \tag{6}$$

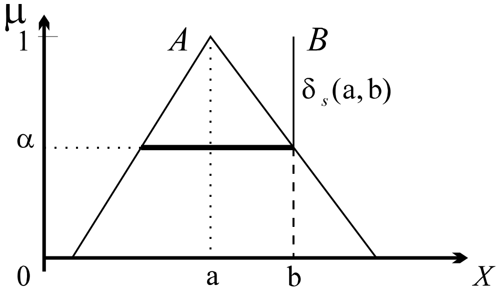


Figure 6

The challenge during the employment of this method is to find approximate scaling functions for both the antecedent and the consequent universes, which give good descriptions in case of non-Ruspini partitions, too. Scaling functions for the case of triangle and trapezoid shaped fuzzy sets are given in [11]. In consequence of the creation of the vague environments of the antecedent and consequent universes, the vague environment of the rule base is established, as well. In this environment each rule is represented by a point. If the observation is a crisp set, the conclusion, which will be crisp, can be also determined employing any interpolative or approximate technique.

The possibility of creation of the antecedent and consequent vague environments in advance ensures the fastness and hereby the applicability of the method for real-time tasks. Thus, only the interpolation of the points describing the rule base has to be made during the functioning of the system. In case of fuzzy observations the antecedent environment should be created taking into consideration the shape of the set, which describes the input.

Figure 7 presents the partitions, the scaling function and the curve built from the points defined by the existent two rules and the points interpolated for the case of a one dimensional antecedent universe supposing crisp observations. The method satisfies the *conditions* 1-4, 5a, 6 and 8.

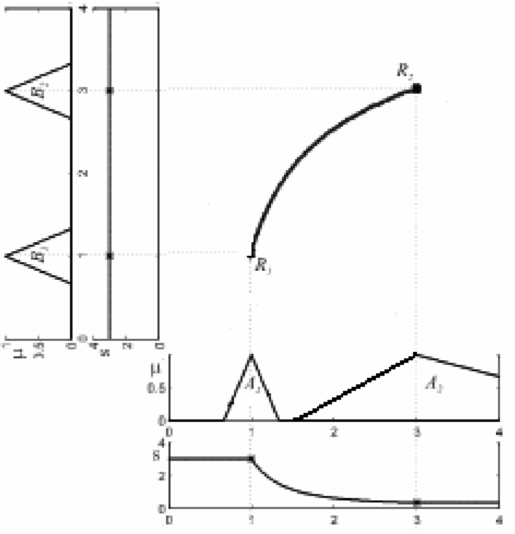


Figure 7

FIVE

3.3 The Generalized Methodology

Baranyi, Kóczy and Gedeon proposed in [1] a generalized methodology for the task of the fuzzy rule interpolation. In the centre of the methodology stands the interpolation of the fuzzy relation. A reference point, which can be identical with e.g. the centre point of the core, is used for the characterization of the position of fuzzy sets. The distance of fuzzy sets is expressed by the distance of their reference points. The interpolation is broken down to two steps.

In the first step an interpolated rule is produced, whose antecedent has at least a partial overlapping with the observation and whose reference point coincides with the reference point of the observation. This task is divided into three stages. First with the help of a set interpolation technique the antecedent of the new rule is produced. Next the reference point of the conclusion is interpolated going out from the position of the reference points of the observation and the reference points of the sets involved in the rules taken into consideration. The applied technique can be a non-linear one, too. Hereupon the consequent set is determined similarly to the antecedent one. Several techniques are suggested in [1] for the task of set interpolation (e.g. SCM, FPL, FVL, IS-I, IS-II). In this paper the solid cutting method is presented in section 3.3.1. If λ_a (Fig. 8) denotes the ratio, in which the reference point of the observation divides the distance between the reference points of the neighbouring sets into two parts and λ_c denotes the similar ratio on the consequent side, the function λ_c=f(λ_a) defines the position of the reference point of the consequent set. Through the selection of the function f() a whole family of linear (λ_c=λ_a) and non-linear interpolation techniques can be derived. This is also a possibility for parameterisation (tuning) of the methodology, which ensures the adaptation to the nature of the modelled system.

The approximated rule is considered as part of the rule base in the second step. The conclusion corresponding to the observation is produced by the help of this rule. As the antecedent of the estimated rule generally does not fit perfectly to the observation, some kind of special single rule reasoning is needed. For example the similarity transfer method introduced in [13], the revision principle based FPL and SRM techniques presented in [24] or the scale and move transformations based method [5] can be applied with success in this step. As a precondition for the revision principle based methods, it should be mentioned that the support of the antecedent set has to coincide with the support of the observation. Generally this is not fulfilled. In such cases the fuzzy relation (rule) obtained in the previous step is transformed first, in order to meet this condition.

Owing to the modular structure of the methodology in both of the steps one can choose from many potential methods if some conventional elements (e.g. distance measure) are used consequently. Based on the analysis in [1] and [22] the methodology can be characterized as follows. *Conditions* 1-4, 5a and 8 are satisfied applying any of the suggested methods in [1]. In case of triangle shaped fuzzy sets the *condition* 7 is also fulfilled by those techniques. *Condition* 9 is also satisfied if SCM or FPL is used in the first step and FPL is used in the second step.

3.3.1 The Solid Cutting Method

The key idea of the solid cutting method (SCM) [2] developed by Baranyi et al. is to define vertical axes at the reference points of the two antecedent sets (A_1 and A_2) which flank the observation (A*) and after that to rotate these sets by 90º around the vertical axes. The virtual space created in such mode is determined by the orthogonal coordinate axes S, X and μ. The rotated sets will be situated in parallel plane to the plane $\mu x S$ (Fig. 8).

In the next step a solid is generated fitting a surface on the contour and support of the sets. After this the solid is cut by the reference point of the observation with a plane parallel with $\mu x S$. Turning back the cross section by 90º one will obtain the antecedent set of the estimated rule. The consequence of the new rule is determined similarly by knowing the two consequent sets and the reference point.

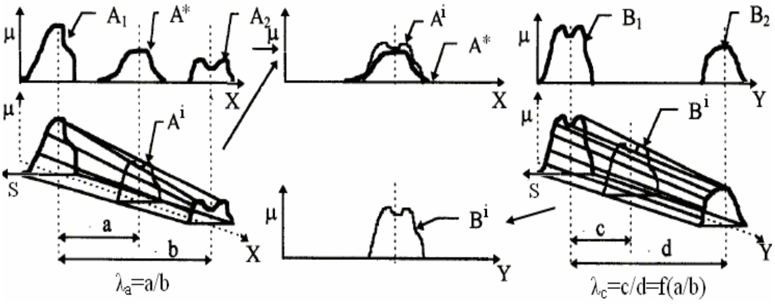


Figure 8
SCM [2]

3.3.2 Single Rule Reasoning Based on Semantic Revision

The antecedent set (A^i) of the interpolated rule generally does not fit perfectly to the observation ($A*$), therefore some kinds of special single rule reasoning techniques (SRRT) are needed in the second step. Further on the key ideas of the semantic revision based methods are reviewed.

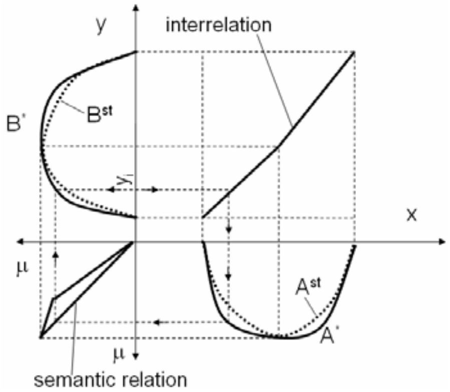


Figure 9

The SRM was introduced by Shen, Ding and Mukaidono [24]. It relies on the concept of the interrelation and semantic relation functions. The *interrelation function* (IR) is a mapping between the elements of two fuzzy sets. It defines which points of the sets are related to each other (Fig. 9 first quarter). The *semantic relation function* (SR) is a mapping between the membership values of the interrelated points of two fuzzy sets (Fig. 9 third quarter). In the Fig. 9 there are two semantic relation functions describing the relation between the left flanks and between the right flanks separately.

As a precondition of the application of the SRM it should be mentioned that the support of the antecedent set has to coincide with the support of the observation and the heights of the rule antecedent and the observation A^* have to be the same. In order to fulfil the conditions generally two transformations of the interpolated fuzzy relation (rule) are necessary. First the technique called *Transformation of the Fuzzy Relation* (TFR) [1] transforms (stretches or shrinks) the interrelation area proportionally by the help of set transformations in order to ensure the needed coincidence of the supports. Secondly the algorithm called *Transformation of the Semantic Relation* [1] modifies the semantic relation area corresponding to the height of A^*.

The SRM methods go out from the transformed sets A^{st}, B^{st} and the transformed relation areas (Fig. 9). Its key idea is the assumption that between A^* and B^* there exists the same interrelation and semantic relation as between A^{st} and B^{st}. It means that substituting A^{st} by A^* and abandoning B^{st} the approximated conclusion can be determined using the existing interrelation and semantic relation. Thus going out from the point y_i that belongs to the left edge of B^* its membership value can be determined following the dashed lines in the directions given by the arrows.

3.4 Interpolation with Generalized Representative Values

The IGRV method proposed by Huang and Shen [5] follows an approach similar to the generalized methodology. In the first phase, a representative value (RV) is determined for each used set. Its function is the same as the function of the reference point in the generalized methodology. It can be calculated by different formulas depending on the demands of the application. The centre of gravity played this role in the first variant of the method [4], which was developed for triangle shaped fuzzy sets. In case of an arbitrary polygonal fuzzy set the weighted average of the x coordinates of the node (break) points is suggested as RV.

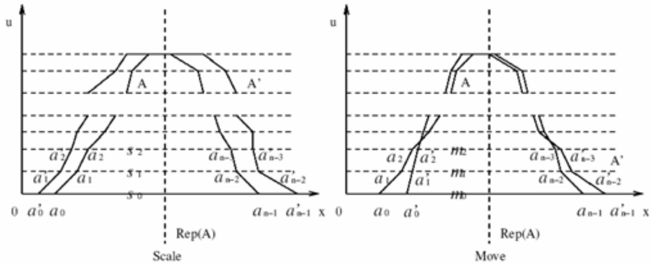


Figure 10
Scale and move transformations [5]

The definition mode of the representative value influences only the position of the estimated rule, but not the shape of the sets involved in the rule. Further on the Euclidean distance of the RVs of the sets are considered as the distance of the sets.

The antecedent of the approximated rule is determined by its α-cuts in such way that two conditions have to be satisfied. First its representative value has to coincide with the RV of the observation. Secondly the endpoints of the α-cuts of the observation have to divide the distance of the respective (left or right) endpoints of the α-cuts of the neighbouring sets in such proportion as the RV of the observation divides the distance of the RVs of these sets. Following the same proportionality principle the representative value and the shape of the consequent of the approximated rule are determined.

In the second phase, the similarity of the observation and the antecedent part of the new rule is characterized by the scale and move transformations needed to transform the antecedent set into the observation. The method was developed primordially for the case of polygonal shaped fuzzy sets. It is applicable in the case of multidimensional antecedent universes, too. In terms of classification, it can be considered as an α-cut based technique, because the scale and move transformation ratios are calculated for each level corresponding to node (break) points of the shape of sets.

The method is well applicable in case of polygonal shaped sets, but the checking and constraint applications done at each α-level for the sake of the conservation of convexity increase the computational complexity of the technique.

It can be tuned at two points. First one can choose the formula for the representative value. Secondly one can choose the method for the calculation of the resulting transformation ratios in the case of multidimensional antecedent universes. On the grounds of the analysis in [5] it can be stated that the method satisfies the *conditions* 1, 2, 3, 4, 5a, 8, and 9.

Conclusions

Systems working with a conventional inference method in case of a sparse rule base cannot produce a result for all the possible input universe values. In such cases the system should adopt an approximate reasoning technique for the estimation of the conclusion. The surveyed methods can be classified into two fundamental groups depending on whether they are producing the result in one or two steps.

In the first part of this paper a general condition set was brought together containing the features, which can be expected from interpolative methods considering an application-oriented viewpoint. After this some well-known methods were surveyed emphasizing their basic ideas, significant characteristics and the conditions they are fulfilling.

References

[1] Baranyi, P., Kóczy, L. T. and Gedeon, T. D.: A Generalized Concept for Fuzzy Rule Interpolation. In IEEE Transaction On Fuzzy Systems, ISSN 1063-6706, Vol. 12, No. 6, 2004, pp. 820-837

[2] Baranyi, P., Kóczy, L. T.: A General and Specialised Solid Cutting Method for Fuzzy Rule Interpolation, In J. BUSEFAL, URA-CNRS, Vol. 66, Toulouse, France, 1996, pp. 13-22

[3] Gedeon, T. D., Kóczy, L. T.: Conservation of fuzziness in the rule interpolation, Intelligent Technologies, International Symposium on New Trends in Control of Large Scale Systems, Vol. 1, Herlany, 1996, pp. 13-19

[4] Huang, Z. H., Shen, Q.: A New Interpolative Reasoning Method Based on Center of Gravity, in Proceedings of the 12th International Conference on Fuzzy Systems, Vol. 1, pp. 25-30, 2003

[5] Huang, Z., Shen, Q: Fuzzy interpolation with generalized representative values, in Proceedings of the UK Workshop on Computational Intelligence, pp. 161-171, 2004

[6] Jenei, S.: Interpolation and Extrapolation of Fuzzy Quantities revisited - (I), An Axiomatic Approach, Soft Computing, ISSN: 1432-7643, 5 (2001), pp. 179-193

[7] Kóczy, L. T., Hirota, K.: Ordering, distance and closeness of fuzzy sets, Fuzzy Sets and Syst., Vol. 59, 1993, pp. 281-293

[8] Kóczy, L. T., Hirota, K.: Rule interpolation by α-level sets in fuzzy approximate reasoning, In J. BUSEFAL, Automne, URA-CNRS, Vol. 46, Toulouse, France, 1991, pp. 115-123

[9] Kóczy, L. T., Kovács, Sz.: Shape of the fuzzy conclusion generated by linear interpolation in trapezoidal fuzzy rule bases, in Proc. 2nd Eur. Congr. Intelligent Techniques and Soft Computing, Aachen, Germany, 1994, pp. 1666-1670

[10] Kóczy, L. T., Kovács, Sz.: The convexity and piecewise linearity of the fuzzy conclusion generated by linear fuzzy rule interpolation, In J. BUSEFAL 60, Automne, URA-CNRS. Toulouse, France, Univ. Paul Sabatier, 1994, pp. 23-29

[11] Kovács, Sz., Kóczy, L. T.: Application of an approximate fuzzy logic controller in an AGV steering system, path tracking and collision avoidance strategy, Fuzzy Set Theory and Applications, Tatra Mountains Mathematical Publications, Mathematical Institute Slovak Academy of Sciences, Vol. 16, pp. 456-467, Bratislava, Slovakia, (1999)

[12] Shepard, D.: A two dimensional interpolation function for irregularly spaced data, Proc. 23rd ACM Internat. Conf., (1968) 517-524

[13] Shi, Y., Mizumoto, M., Wu, Z. Q.: Reasoning conditions on Kóczy's interpolative reasoning method in sparse fuzzy rule bases, Fuzzy Sets Syst., Vol. 75, pp. 63-71, 1995

[14] Tikk, D., Baranyi, P., Gedeon, T. D., Muresan, L.: Generalization of the Rule Interpolation method Resulting Always in Acceptable Conclusion, Bratislava, Slovakia, Tatra Mountains, Math. Inst. Slovak Acad. Sci., 2001, Vol. 21, pp. 73-91

[15] Tikk, D., Baranyi, P., Yam, Y., Kóczy, L. T.: Stability of a new interpolation method, in Proc. IEEE Conf. Syst. Man. and Cybern. (SMC '99), Tokyo, Japan, 1999, pp. III/7-III/9

[16] Tikk, D., Baranyi, P.: Comprehensive analysis of a new fuzzy rule interpolation method, IEEE Trans Fuzzy Syst., Vol. 8, June 2000, pp. 281-296

[17] Tikk, D., Joó, I., Kóczy, L. T., Várlaki, P., Moser, B., Gedeon, T. D.: Stability of interpolative fuzzy KH-controllers. Fuzzy Sets and Systems, 125(1) pp. 105-119, January 2002

[18] Tikk, D.: Investigation of fuzzy rule interpolation techniques and the universal approximation property of fuzzy controllers, Ph. D. dissertation, TU Budapest, Budapest, 1999

[19] Vass, Gy., Kalmár, L., Kóczy, L. T.: Extension of the fuzzy rule interpolation method, in Proc. Int. Conf. Fuzzy Sets Theory Applications (FSTA '92), Liptovsky M., Czechoslovakia, 1992, pp. 1-6

[20] Wong, K. W., Gedeon, T. D., Tikk, D.: An improved multidimensional α-cut based fuzzy interpolation technique, in Proc. Int. Conf Artificial Intelligence in Science and Technology (AISAT'2000), Hobart, Australia, 2000, pp. 29-32

[21] Yam, Y., Kóczy, L. T.: Representing membership functions as points in high dimensional spaces for fuzzy interpolation and extrapolation. Technical Report CUHK-MAE-97-03, Dept. Mech. Automat. Eng., The Chinese Univ. Hong Kong, Hong Kong, 1997

[22] Mizik, S.: Fuzzy Rule Interpolation Techniques in Comparison, MFT Periodika 2001-04, Hungarian Society of IFSA, Hungary, 2001, http://www.mft.hu

[23] Wong, K. W., Tikk, D., Gedeon, T., Kóczy, L. T.: Fuzzy Rule Interpolation for Multidimensional Input Spaces With Applications: A Case Study, IEEE Transactions On Fuzzy Systems Vol. 13, No. 6, Dec. 2005, pp. 809-819

[24] Shen, Z., Ding, L., Mukaidono, M.: Methods of revision principle, in Proc. 5th IFSA World Congr., 1993, pp. 246-249

A Canonical Form of RT-Level FSM Controlled Data Path Descriptions for Formal Verification

Péter Keresztes

Széchenyi István University
Egyetem tér 1, H-9026 Győr, Hungary
keresztp@sze.hu

Abstract: The paper proposes a new canonical form for RT-level descriptions, which can be systematically generated from both the specification and the structural description. The verification can be executed with the comparison of the two generated canonical form descriptions.

1 Introduction

When it comes to the designing of digital systems, a description in accordance with a well-chosen canonical form provides grounds for the efficient methods of the formal verification and the symbolic simulation, alike. The logic (gate-level) synthesis, along with the verification and the symbolic simulation are all based on the canonical forms, which borrows its tools from the classic switching algebra. In the aspect of their application on computer design systems, particularly successful was Roth's cube algebra, which is based on a new wording of Boole's canonical forms [1].

The descriptions of the register transfer level have up to the present lacked the universality and heuristic power, which characterises the switching algebra. Thus, the canonical forms employed on the register level could only be applied to a restricted scale of tasks. To this category belongs, for instance, the Taylor-polynomial method, which is capable of verifying the register-level structures of arithmetic expressions, but has its limits within this very class [2], [3].

The implementation of the register transfer level canonical description suggested by the author of the present paper is conditional on the same requirements as those forming the principle of the most part of designing methods. The data-path structure is controlled by a synchronous finite state machine (FSM), as a controller built around a core. The structure must clearly reflect that in a specific state of the FSM, as an interval:

1 Which sub-paths of the data-path are switched active by the multiplexers,

and

2 Into which registers and on what conditions occurs entering of data.

On condition that the structure's description meets the requirements above, the canonical form, as suggested by this paper, can be prepared.

At the same time, an identical canonical description is gained from the algorithm-level specification, which is a behavioural description, formulated in one of the high level programming languages. If the canonical description, gained from the structure, and the behavioural description are provably homomorphous, – even at the expense of certain permissible transformations – the verification process can be considered successful.

2 Decomposition of Sequential Behavioral Descriptions

We decompose the program, constituted by sequential statements, into a hierarchical structure of modules, between the statements modifying the control, as bordering points. In the sequential subset of VHDL-processes the control branch statements are the following:

begin . . . . end

wait until . . . .

for . . loop. . . . .end loop

while . . .loop . . . end loop

if . . . then . . .else . . . end if

The example below is the abstract style VHDL behavioural description of a hardware unit in charge of carrying out the algorithm of square root calculation. *Figure 1* shows the way we decompose the description into modules, and the way these modules and their attachments constitute the state-graph of an abstract state machine. It is important to formulate the variable-assignment statements of the description through functions that are implemented by the components (function-units) of the hardware structure.

```
library work; use work.sqrtpack.all;
entity SQRT_UNIT is
  port ( START : in bit;
         READY : inout bit := '1';
         RESET : in bit;
         pe : in real := 0.0;
         px : in real:= 0.0;
         py : inout real := 0.0;
         ph1, ph2 : in bit);
end SQRT_UNIT;

architecture BEH of SQRT_UNIT is
 begin
 process
   variable e, x, y, cy, ny, v : real := 0.0;
   variable d : real := 1.0;
   variable f : bit := '1';
   variable g : bit;
  begin
   wait until START = '1';
   READY <= '0';
   wait for 1 ns;
   e := pe; x := px;
   cy := Fi(x);
   wait for 1 ns;

while f = '1' loop
      v := MD(div, x, cy);
      v := AS(add, cy, v);
      ny := MD(mult, 0.5, v) ;
      d := AS(sub,ny,cy);
      g := Cm(d, 0.0);
      if g = '0' then
         d := AS(sub, 0.0, d);
      end if;
      cy := ny;
      f := Cm(d, e);
    end loop;
   wait for 1 ns;
    py <= cy;
    READY <= '1';
end process;
end BEH;
```

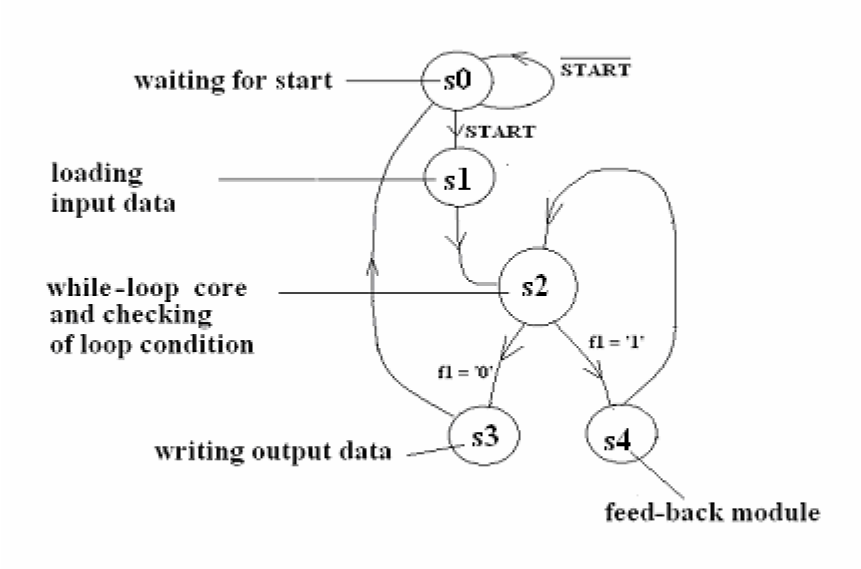


Figure 1

The decomposition of the square root algorithm into sequential modules

3 Generating Value-target Event-driven Data-flow Blocks from Behavioural Description

Consider the variable-assignment statements of a sequential module and the values ordered to the variables **v1, v2, . . . vj . . . .vn** by the sequence. Pick out the value of **vj** next in line, resulting from the next-in-line variable assignment. Formulate this in the following substitution expression:

$$vj(p+1)\ = E[\ldots v_k / vk(p) \ldots]$$

A value next in line of variable $\mathbf{v_j}$ can be calculated through the substitution of the present values of the variables of the right hand side into the variable-assignment statement. If we number the values of the variables of the sequence, from *v1(0), v2(0), . . . vj(0), . . .vn(0)* up to those terminal values of maximum indexes *v1(t), v2(t), . . . vj(t), . . .vn(t)*, ordering one target to each and every value of each and every variable, and on the other hand, we order to each variable-assignment an event-driven concurrent statement,

$$wj(p+1)\ \texttt{<=}\ E\,[\ldots vk / wk(p) \ldots]$$

then from these statements we attain an event-driven dataflow-block, which can be ordered to the sub-sequence. This block is termed the value-target block (**VTB**) of the sequential module.

The **VTB** at rest is

$$wj\,(t) = E\,[\,.\,.\,.vk\,/\,wk(t)\,.\,.\,.\,]$$

It is conceivable that if the initial value of the variable v_j is equal to the initial value of the target w_j, ordered to it, then the value of v_j, with which it leaves the sequence module, is also equal to the terminal value of the target of the maximal index. One sequence is therefore *value-equivalent* to the value-tracking block gained from it. See a simple example:

```
SEQ1: begin
          for i in 1 to 4 loop
            a := a + 1;
        end loop;
end;
```

```
VTB1 : block
    begin
      a1 <= a0 + 1;
      a2 <= a1 + 1;
      a3 <= a2 + 1;
      a4 <= a3 + 1;
end block;
```

A more complex one:

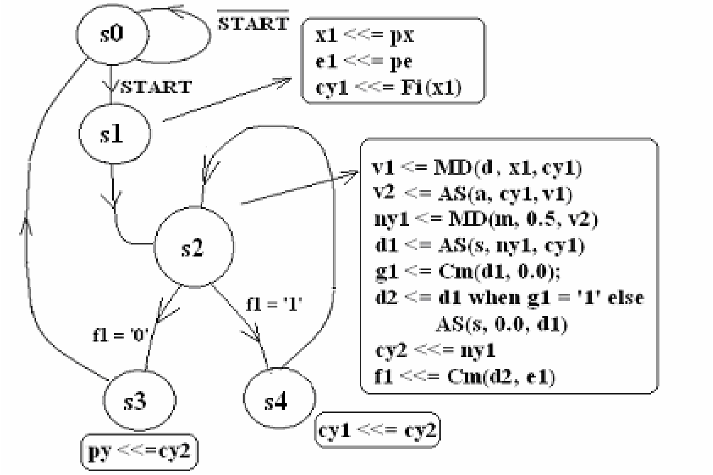
```
SEQ2 : begin
              if e < 0 then a := b * c;
                            d := a + b;
              elsif  e = 0 then
                             a := b;
              else
                             d := a;
              end if;
        end;
```

```
VTB2 :  block  begin
                  a1 <= b0 * c0  when
                              e0 < 0  else
                  b0 when e0 = 0  else
                              a0;
                  d1 <= a1 + b0 when
                              e0 < 0 else
                  d0 when e0 = 0  else
                              a0;
          end block;
```

Now complement the abstract state-transition graph, attained from the square root algorithm, with the **VTB**s of the particular modules. Hence will be obtained the description in accordance with *Figure 2*. Hereafter, this is regarded as the canonical form of the specification.

4 Notations

In *Figure 2* a possible form of FSM controlled value-target blocks are shown. The meaning af notations which are used in the blocks can be explained by the semantics of VHDL statements. *Table 1* shows that form of canonical description of the specification, which will be compared with the canonical form of the structure.



```
x1 <<= px
e1 <<= pe
cy1 <<= Fi(x1)
```

```
v1 <= MD(d, x1, cy1)
v2  <= AS(a, cy1, v1)
ny1 <= MD(m, 0.5, v2)
d1 <= AS(s, ny1, cy1)
g1 <= Cm(d1, 0.0);
d2 <= d1 when g1 = '1' else
        AS(s, 0.0, d1)
cy2 <<= ny1
f1 <<= Cm(d2, e1)
```

```
cy1 <<= cy2
```

```
py <<=cy2
```

Figure 2

The canonical form without time-refinement of the square root calculation's specification

Target/state	**s0**	**s1**	**s2**	**s3**	**s4**
v1			**MD(d,x1,cy1)**		
ny1			**MD(m,0.5,py)**		
v2			**AS(a,cy1,v1)**		
d1			**AS(s,ny1,cy1)**		
g1			**Cm(d1, 0.0)**		
d2			**d1 when g1 = ’1’ else AS(s, 0.0, d1)**		
e1		**pe**			
x1		**px**			
cy1		**Fi(x1)**			**cy2**
cy2			**ny1**		
f1			**Cm(d2, e1)**		
py				**cy2**	

Table 1

Canonical form of specification. The simple- (<=) and the register-type (<<=) transactions are isolated parts of the first column.

The simple transaction **v1 <= MD(d, x1, y1)** given in a box ordered to state **s2** of **FSM** can be expressed as follows:

```
    v1 <= MD(d, x1, y1) when fsm_state = s1 else anyvalue;
```

The transaction for **v1** is a *non-register-type* statement.

An other transaction, for example **cy1 <<= Fi(x1)** given in the fsm-state **s1** together with the other transaction with the same target in fsm-state **s4**, (**cy1 <<= cy2**) can be interpeted as follows:

```
   cy1 <= Fi(x1) when fsm_state = s1 and fsm_phase = ph2 else
              cy2 when  fsm_state = s4 and fsm_phase = ph2 else
              cy1;
```

The two transactions for **cy1** constitute *register-type* statement. It has to be emphasided that each register type transaction is executed by a *phase* of an **FSM** state. The phase has to be an inner time interval of the of the **FSM** state.

5 Characteristics of the Proposed Canonical Form Description

Two sequential descriptions can be fully equivalent in spite of the number of variables or the order of statements within them being different. The canonical form described above shows some very important features. These are the following:

1 Unaffected by the number of the variables of the specification's equivalent forms.

2 Unaffected by the order of statements in the modules' equivalent forms.

3 Unaffected by the number of FSM states deriving from hardware limitations.

4 Unaffected by the allocations of function-unit, register and multiplexer, which derive from hardware limitations.

The first characteristic derives from the fact that the description orders the target-signals to the values of the variables, which means that the canonical forms of two equivalent sequential modules are identical, irrespective of the difference between the number of their respective variables. The second feature derives from the fact that we convert the modules into data-flow blocks composed of concurrent statements, and thus the canonical forms of equivalent sequential modules applying different orders of statements are also identical. The third characteristic derives from the fact that the states, whose number has been increased because of the necessity generated by the hardware limitations, can be contracted during the transformations of the canonical form that describes the structure. The fact that units lose their identities during the transformations of the structure-describing canonical form, and appear in the changed canonical description only through their functions (similarly to the way they do in the canonical description of the specification) accounts for the fourth characteristic.

6 Process of Verification of a RT-level Unit

The **RT**-level structure to be verified is shown in *Figure 3*, *Figure 4*, and *Table 2*. The description has to contain the structure of **DATA-PATH** (*Figure 3*) and the state-transition graph of the **FSM** (*Figure 4*). The fuction units of the **DATA-PATH**:

- One multiplier/divider unit (**M/D**)
- Two adder/subtractor units (**A/S**)
- Two comparators (**Cm**)
- A special look-up table unit for deriving of initial approximation of square-root. (**Fi**)

Above these components the **DATA-PATH** contains 6 registers and 7 multiplexers. *Table 2* shows the initial form of the structural description to be verified. There is a part of the targets which contain outputs of multiplexers and function units, while another part of them contain outputs of registers. These parts are isolated in the left side of the list. There are state-independent transactions in the structure, and they are isolated in the left side of the list.

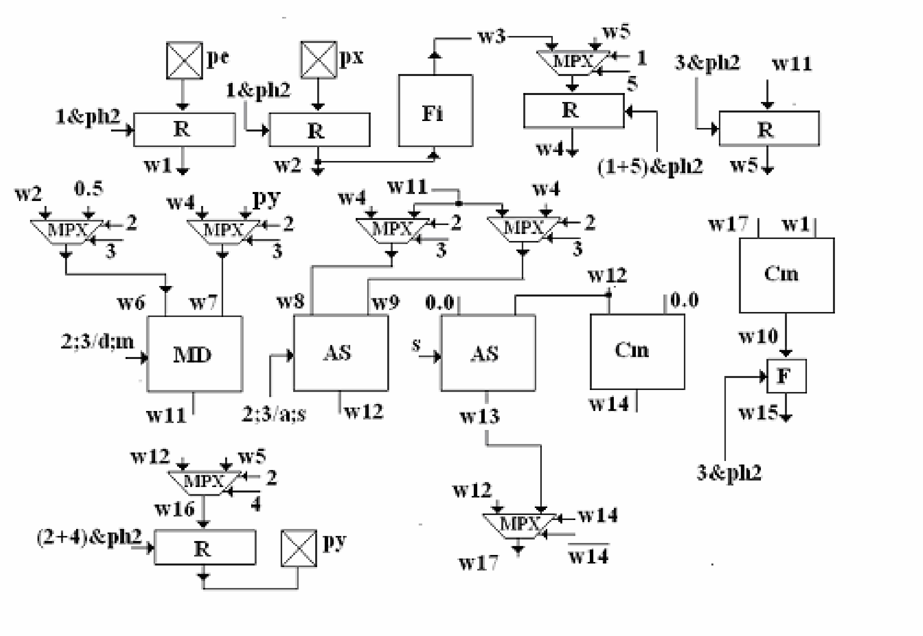


Figure 3
RT-level structure of square-root calculation unit

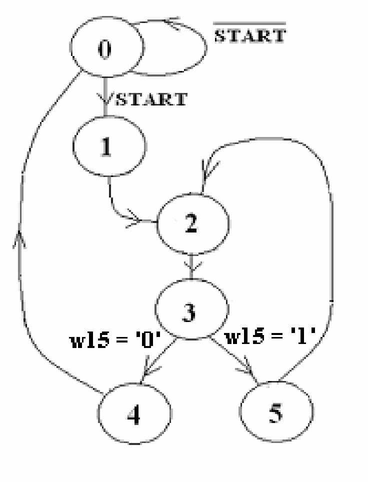


Figure 4
The graf of counter-based FSM, which controls the data-path of square-root calculation unit

Target/state	0	1	2	3	4	5
w3	Fi(w2)					
w6			w2	0.5		
w7			w4	py		
w8			w4	w11		
w9			w11	w4		
w10	Cm(w17,w1)					
w11			MD(d,w6,w7)	MD(m,w6,w7)		
w12			AS(a,w8,w9)	AS(s,w8,w9)		
w13	AS(s,0.0,w12)					
w14	Cm(12, 0.0)					
w16			w12		w5	
w17	w12 when w14 = '1' else w13					
w1		pe				
w2		px				
w4		w3				w5
w5				w11		
w15				w10		
py			w16		w16	

Table 2

The source of the structural canonical description, which is derived from the implementation

7 Transformation Steps of the Verification Process

The application of the following transformation steps leads to the canonical form of structural description which can be compared with the canonical form of the specification. They are based on the semantical equivalency of some parts of the structural description and corresponding abstract data-flow expressions. The name of each step is a reference to the structural analogy of the given transformation.

7.1 Placement of State-Independent Transactions into States

The first step of transformation is the placement of the state-independent transactions into those states, in which the target of the transaction, or the target of another transaction which is driven by it is stored in a register. This step is based on the recognition that the target-value of a state-independent transaction is *'don't care'* in those states, in which it is not stored.

For example, given a state independent transaction

wi <= EXPi

and **wi** is used in state **nl** as follows:

nl : wj <= EXPj(. . . wi . . .), wk <<= wj.

In this case the result of the placement is the following:

S[nl] : wi <=EXPi wj <= EXPj(. . . wi . . .) wk <<= wj

7.2 Node Elimination

In the second phase of the transformations those targets are eliminated inside a given **FSM** state, which are not stored in the given state, and they are used at the right side of another target. It can be shown, that this step can eliminate all the nodes, the signals represented by which do not belong to a behavioural description. Assume that in fsm-state **nl** the following transactions are given:

nl : wi <= wj wk <= EXP(. . . wi . . .)

The result of the node elimination is as follws:

S[nl] : wk <= EXP(. . . wj . . .)

7.3 Merging Subsequent Loopless States

The number of **FSM** states in canonical specifications is minimum, but in the structural description because of the hardware constraints it can be much higher. The **FSM** states that are introduced only because of the contraints of the number of function units are subsequent, and there is no control feedback between them.

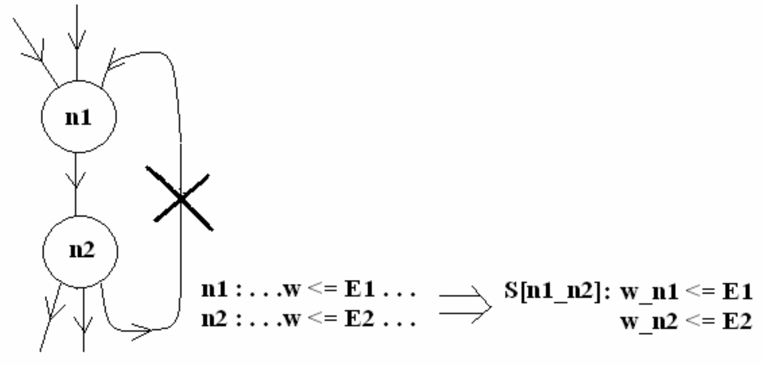


Figure 5

Merging subsequent FSM states with data-independent transactions

To get closer to the canonical form, a merging of these states is proposed. *Figure 5* shows the simpler case, when there is no such signal which is stored in state **n1** and used in state **n2**. In a more complex case, when a value of a signal is stored in a register, and it is used in the next state, the register after merging is eliminated. (*Figure 6*)

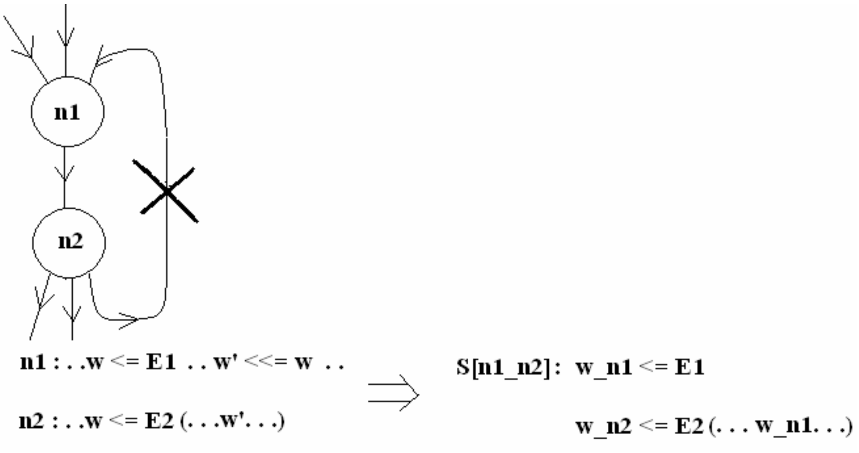

n1
n2
n1 : . .w <= E1 . . w' <<= w . .
n2 : . .w <= E2 (. . .w'. . .)
S[n1_n2]: w_n1 <= E1
w_n2 <= E2 (. . . w_n1. . .)

Figure 6

Merging subsequent FSM states with a common signal, which is stored in state **n1** for state **n2**

8 Generation of the Structural Canonical Form of Square-Root Unit

The following series of tables from *Table 3* to *Table 7* illustrates the verification flow of the hardware implementation of square-root procedure. The intermediate forms and the application of the three transformation steps lead to the structural canonical description.

Table 3 is the result of placement of state-independent transactions. For example the transaction **w3 <= Fi(w2)** was placed in **state 1**, because **w3** is stored in **ph2** phase of the **state 1**. The result of node eliminations is shown in *Table 4*. For example **w6** is eliminated, since **w6** is driven by **w2** in **state 2**, and **w6** is used in the driver **MD(d, w6, w7)** in the same state. So **w2** substitutes **w6** in driver of **w11**.

The result of merging state **'2'** and state **'3'** are shown in the *Table 5*. The *Figure 5* which shows the state-transition graf of the implemntation, proofs that '**2**' and '**3**' are subsequent and loopless states. Since the targets **w11** and **w12** are driven in both states, after the merging both of them have to duplicated.

Target/state	0	1	2	3	4	5
w3		Fi(w2)				
w6			w2	0.5		
w7			w4	py		
w8			w4	w11		
w9			w11	w4		
w10				Cm(w17,w1)		
w11			MD(d,w6,w7)	MD(m,w6,w7)		
w12			AS(a,w8,w9)	AS(s,w8,w9)		
w13				AS(s,0.0,w12)		
w14				Cm(12, 0.0)		
w16			w12		w5	
w17				w12 when w14 = ’1’ else w13		
w1		pe				
w2		px				
w4		w3				w5
w5				w11		
w15				w10		
py			w16		w16	

Table 3

Result o the placement of the state independent transactions

Target/state	0	1	2	3	4	5
w11			MD(d, w2, w4)	MD(m, 0.5, py)		
w12			AS(a, w4, w11)	AS(s, w11, w4)		
w14				Cm(12, 0.0)		
w17				w12 when w14 = ’1’ else AS(s, 0.0, w12)		
w1		pe				
w2		px				
w4		Fi(w2)				w5
w5				w11		
w15				Cm(w17,w1)		
py			w12		w5	

Table 4

Result of the node elimination

Table 6 is the result of the attempt, the goal of which to find a consistent cross-reference list between the nodes of the canonical description of the structure and the signals of the canonical specification. If the nodes of *Table 5* are replaced by the signals of the cross-regference list, *Table 7* is derived. It is obvious, that the structural canonical description covers the canonical specification, because the corresponding signals appear in every correspondig cells of the table, where the driver is specified.

Target/state	**0**	**1**	**2 3**	**4**	**5**
w11_1			**MD(d, w2, w4)**		
w11_2			**MD(m, 0.5, py)**		
w12_1			**AS(a, w4, w11_1)**		
w12_2			**AS(s, w11_2, w4)**		
w14			**Cm(w12_2, 0.0)**		
w17			**w12_2 when w14 = '1' else** **AS(s, 0.0, w12_2)**		
w1		**pe**			
w2		**px**			
w4		**Fi(w2)**			**w5**
w5			**w11_2**		
w15			**Cm(w17,w1)**		
py			**w12_1**	**w5**	

Table 5
Result of the merging state '2' and state'3'

Signals from the structural desription	**Signals from the canonical behavioural description**
w1	**e1**
w2	**x1**
w4	**cy1**
w11_1	**v1**
w12_1	**v2**
w17	**d2**
w12_2	**d1**
w11_2	**ny1**
w5	**cy2**
w15	**f1**
w14	**g1**

Table 6
Equivalence between the signals of the structure and the signals of specification

Target/state	**s0**	**s1**	**s2**	**s3**	**s4**
v1			**MD(d,x1,cy1)**		
ny1			**MD(m,0.5,py)**		
v2			**AS(a,cy1,v1)**		
d1			**AS(s,ny1,cy1)**		
g1			**Cm(d1, 0.0)**		
d2			**d1when g1 = ’1’ else AS(s,0.0,d1)**		
e1		**pe**			
x1		**px**			
cy1		**Fi(x1)**			**cy2**
cy2			**ny1**		
f1			**Cm(d2,e1)**		
py			**v2**	**cy2**	

Table 7

Application af signal equivalences as a final step of verification

Conclusions

The new canonical form detailed above seems to be capable of developing an algorithm and an automatic verification system. The work intended to work out an implementation of the algorithm has been started.

References

[1] M. A. Breuer: Design Automation of Digital Systems, Prentice-Hall Inc, 1972

[2] M. Ciesielsky, P. Kalla, Z. Zeng and B. Rouzeyre: Taylor Expansion Diagrams: A new Representation for RTL Verification, IEEE Intl. High Level Design Validation and Test Workshop (HLDVT’01), 2001, pp 70-75

[3] P. Kalla, M. Ciesielsky, E. Boutillon, E. Martin: High Level Design Verification Using Taylor Expansion Diagrams: First Results, IEEE Intl. High Level Design Validation and Test Workshop (HLDVT’02), 2002, pp 13-17

Product Type Operations between Fuzzy Numbers and their Applications in Geology

Barnabás Bede

Department of Mechanical and System Engineering, Bánki Donát Faculty of Mechanical Engineering, Budapest Tech Polytechnical Institution
Népszínház u. 8, H-1081 Budapest, Hungary
e-mail: bede.barna@bgk.bmf.hu

János Fodor

John von Neumann Faculty of Informatics, Budapest Tech Polytechnical Institution
Bécsi út 96/B, H-1034 Budapest, Hungary
e-mail: fodor@bmf.hu

Abstract: Multiplicative operations for fuzzy numbers raise several problems both from the theoretical and practical point of view in fuzzy arithmetic. The multiplication based on Zadeh's extension principle and its triangular and trapezoidal approximation is used in several recent works in applications in geology. Recently, new product-type operation are introduced and studied, as e.g. the cross product of fuzzy numbers and the product obtained by the best trapezoidal approximation preserving the expeted interval. We present a comparative study of the above mentioned multiplications with respect to geological applications.

Keywords: fuzzy number, cross product, trapezoidal approximation, geological resource estimation

1 Introduction

Uncertainties in different scientific areas arise mainly from the lack of human knowledge. In many practical situations the uncertainties are not of statistical type. This situation occurs mainly in the case of modeling the linguistic expressions because of their dependence on human judgement. Also, as it is shown in several recent works (see e.g. [1]) the uncertainty on different measurements due to finite resulution of measuring instruments is in many cases more possibilistic than probabilistic, since, in many applications the measurements cannot be repeated. This situation occurs mainly in Geosciences, since in this case we cannot have two holes in the same place in order to repeat the measurement so every experiment can be considered as unique. This shows us that uncertainties on the measurements in geological data are more of possibilistic type than of probabilistic type, since in order to obtain the statistical distribution of a variable we need several experiments. Fuzzy numbers allow us to modell in an easy way these non-probabilistic uncertainties. This justifies the increasing interest on theoretical and practical aspects of fuzzy arithmetic in the last years, especially directed to: operations over fuzzy numbers and properties, ranking of fuzzy numbers and canonical representation of fuzzy numbers.

Usually, the definition of addition and multiplication of fuzzy numbers are based on the extension principle ([15]). A main disadvantage of the multiplicative operation in this case is that by multiplication the shape of $L - R$ type fuzzy numbers (so triangular or trapezoidal numbers) is not preserved. In many situations this problem is solved by approximating the result of the extension principle-based multiplication by a triangular or trapezoidal number. This can lead to unexpected results in the case of iterative application of these appeoximations (i.e. can increase or decrease a defuzzified vale in a considerable way). For example in [1] the result of the product obtained by using the extension principle is approximated by a trapezoidal number considering the endpoints of the core and support.

Nevertheless, there exist other directions of development of fuzzy arithmetic. For example, in [11] new operations on fuzzy numbers are defined starting from a representation of a fuzzy number by a location index number and two fuzziness index functions.

Recently, in [2] a new multiplicative operation of product type is introduced, the so-called cross-product of fuzzy numbers and its algebraic and analytic properties are studied. The main point is that this product preserves the shape of $L - R$ fuzzy numbers under multiplication, is consistent to the classical error theory and has good algebraic and metric properties. The idea to define the cross product started from the approximation formulas ([5], p. 55) of the multiplication (obtained by Zadeh extension principle) of two $L - R$ type fuzzy numbers by an $L - R$ type fuzzy number if the spreads are small compared with the means of the numbers. The consistency of the cross product with the classical error theory is also proved in [2].

The above mentioned properties motivate to use the cross product in geological applications as a possible altenative of the product obtained by Zadeh's extension principle, mainly in the case of iterative calculations using products in each iteration.

Recently, in [9] a new method is proposed for the approximation of a fuzzy number by a trapezoidal number. This method is the best in some sense and the properties shown in [9] motivate the use of this method in order to approximate the extension priciple-based product by a trapezoidal number. This operation (regarded as a new product) can be also useful in several geological applications.

In Section 2 we recall some concepts from fuzzy arithmetic. In Section 3, we summarise the definition and some properties of the cross product. In Section 4, we discuss the method introduced in [9]. In Section 5 we propose a practical, comparative study of the previously discussed product-type operations. Applications in Geology discussed in this section concern nuclear safety assessment of a repository and the estimation of the quantity of bauxite in Halimba. At the end of the paper we present some conclusions and further research topics.

2 Basic Concepts

Let us recall the following well-known definition of a fuzzy number. The addition of fuzzy numbers and multiplication of a fuzzy number by a crisp number are provided by Zadeh's extension principle.

Definition 1 A fuzzy number is a function $u : \mathbf{R} \to [0,1]$ with the following properties:

(i) u is normal, i.e., there exists $x_0 \in \mathbf{R}$ such that $u(x_0) = 1$;

(ii) $u(\lambda x + (1-\lambda) y) \geq \min\{u(x), u(y)\}, \forall x, y \in R, \forall \lambda \in [0,1]$;

(iii) u is upper semicontinuous on $\mathbf{R}$, i.e., $\forall x_0 \in \mathbf{R}$ and $\forall \varepsilon > 0$ there exists a neighborhood $V(x_0)$ such that $u(x) \leq u(x_0) + \varepsilon, \forall x \in V(x_0)$;

(iv) The set $\overline{\text{supp}(u)}$ is compact in $\mathbf{R}$, where supp $(u) = \{x \in R; u(x) > 0\}$.

We denote by R_F the set of all fuzzy numbers.

Let $a,b,c \in \mathbf{R}, a < b < c$. The fuzzy number $u : \mathbf{R} \to [0,1]$ denoted by (a,b,c) and defined by $u(x) = 0$ if $x \leq a$ or $x \geq c, u(x) = \frac{x-a}{b-a}$ if $x \in [a,b]$ and $u(x) = \frac{c-x}{c-b}$ if $x \in [b,c]$ is called a triangular fuzzy number.

For $0 < r \leq 1$ and $u \in R_F$ we denote $[u]^r = \{x \in \mathbf{R}; u(x) \geq r\}$ and $[u]^0 = \overline{\{x \in \mathbf{R}; u(x) > 0\}}$. It is well-known that for each $r \in [0,1], [u]^r$ is a bounded closed interval, $[u]^r = \left[\underline{u}^r, \overline{u}^r\right]$. Let $u, v \in R_F$ and $\lambda \in \mathbf{R}$. We define the sum $u+v$ and the scalar multiplication λu by

$$[u+v]^r = [u]^r + [v]^r = \left[\underline{u}^r + \underline{v}^r, \overline{u}^r + \overline{v}^r\right]$$

and

$$[\lambda u]^r = \lambda [u]^r = \begin{cases} \left[\lambda \underline{u}^r, \lambda \overline{u}^r\right], & \text{if } \lambda \geq 0, \\ \left[\lambda \overline{u}^r, \lambda \underline{u}^r\right], & \text{if } \lambda < 0, \end{cases}$$

respectively, for every $r \in [0,1]$.

We denote by $-u = (-1)u \in R_F$ the symmetric of $u \in R_F$.

The product $u \cdot v$ of fuzzy numbers u and v, based on Zadeh's extension principle, is defined by

$$\underline{(u \cdot v)}^r = \min\{\underline{u}^r \underline{v}^r, \underline{u}^r \overline{v}^r, \overline{u}^r \underline{v}^r, \overline{u}^r \overline{v}^r\}$$
$$\overline{(u \cdot v)}^r = \max\{\underline{u}^r \underline{v}^r, \underline{u}^r \overline{v}^r, \overline{u}^r \underline{v}^r, \overline{u}^r \overline{v}^r\}.$$

Surely, the above formulas are not very practical from the computational point of view. Also, let us remark that usually the fuzzy numbers which are used in practical applications are trapezoidal. So, the requirement that a product operation should be shape-preserving seems to be natural.

For this aim, for applications in geology, the following (we will call it old) trapezoidal approximation of the product is used: the endpoints of the 0 and 1 level sets of the product determine a trapezoidal number, which is then regarded as the result of the multiplication (see [1]).

Definition 2 A fuzzy number $u \in R_F$ is said to be positive if $\underline{u}^1 \geq 0$, strict positive if $\underline{u}^1 > 0$, negative if $\overline{u}^1 \leq 0$ and strict negative if $\overline{u}^1 < 0$. We say that u and v have the same sign if they are both positive or both negative.

Let $u, v \in R_F$. We say that $u \prec v$ if $\underline{u}^r \leq \underline{v}^r$ and $\overline{u}^r \leq \overline{v}^r$ for all $r \in [0,1]$. We say that u and v are on the same side of 0 if $u \prec 0$ and $v \prec 0$ or $0 \prec u$ and $0 \prec v$.

Remark 1 If u is positive (negative) then $-u$ is negative (positive).

Definition 3 For arbitrary fuzzy numbers u and v the quantity

$$D(u,v)=\sup_{0\le r\le 1}\{\max\{|\underline{u}^r-\underline{v}^r|,|\overline{u}^r-\overline{v}^r|\}\}$$

is called the (Hausdorff) distance between u and v.

It is well-known (see e.g. [14]) that (R_F,D) is a complete metric space and D verifies $D(ku,kv)=|k|D(u,v)$, $\forall u,v\in R_F$, $\forall k\in R_F$.

3 The Cross Product

In this section we study the theoretical properties of the cross product of fuzzy numbers. Let $R_F^{\ *}=\{u\in R_F\ :\ u$ is positive or negative$\}$. Firstly we begin with a theorem which was obtained by using the stacking theorem ([12]).

Theorem 1 *If* u *and* v *are positive fuzzy numbers then* $w=u\odot v$ *defined by* $[w]^r=\left[\underline{w}^r,\overline{w}^r\right]$, *where* $\underline{w}^r=\underline{u}^r\underline{v}^1+\underline{u}^1\underline{v}^r-\underline{u}^1\underline{v}^1$ *and* $\overline{w}^r=\overline{u}^r\overline{v}^1+\overline{u}^1\overline{v}^r-\overline{u}^1\overline{v}^1$, *for every* $r\in[0,1]$, *is a positive fuzzy number.*

Corollary 1 *Let* u *and* v *be two fuzzy numbers.*

(i) *If* u *is positive and* v *is negative then* $u\odot v=-(u\odot(-v))$ *is a negative fuzzy number;*

(ii) *If* u *is negative and* v *is positive then* $u\odot v=-((-u)\odot v)$ *is a negative fuzzy number;*

(iii) *If* u *and* v *are negative then* $u\odot v=(-u)\odot(-v)$ *is a positive fuzzy number.*

Definition 4 The binary operation $\odot$ on R_F^* introduced by Theorem 1 and Corollary 1 is called cross product of fuzzy numbers.

Remark 1) The cross product is defined for any fuzzy numbers in

$R_F^{\wedge}=\{u\in R_F^*$; there exists an unique $x_0\in\mathbf{R}$ such that $u(x_0)=1\}$, so implicitly for any triangular fuzzy number. In fact, the cross product is defined for any fuzzy number in the sense proposed in [6] (see also [14]).

2) The below formulas of calculus can be easily proved ($r\in[0,1]$):

$$\underline{(u \odot v)}^r = \overline{u}^r \underline{v}^1 + \overline{u}^1 \underline{v}^r - \overline{u}^1 \underline{v}^1,$$

$$\overline{(u \odot v)}^r = \underline{u}^r \overline{v}^1 + \underline{u}^1 \overline{v}^r - \underline{u}^1 \overline{v}^1$$

if u is positive and v is negative,

$$\underline{(u \odot v)}^r = \underline{u}^r \overline{v}^1 + \underline{u}^1 \overline{v}^r - \underline{u}^1 \overline{v}^1,$$

$$\overline{(u \odot v)}^r = \overline{u}^r \underline{v}^1 + \overline{u}^1 \underline{v}^r - \overline{u}^1 \underline{v}^1$$

if u is negative and v is positive. In the last possibility, if u and v are negative then

$$\underline{(u \odot v)}^r = \overline{u}^r \overline{v}^1 + \overline{u}^1 \overline{v}^r - \overline{u}^1 \overline{v}^1,$$

$$\overline{(u \odot v)}^r = \underline{u}^r \underline{v}^1 + \underline{u}^1 \underline{v}^r - \underline{u}^1 \underline{v}^1.$$

3) The cross product extends the scalar multiplication of fuzzy numbers. Indeed, if one of operands is the real number k identified with its characteristic function then $\underline{k}^r = \overline{k}^r = k, \forall r \in [0,1]$ and following the above formulas of calculus we get the result.

The main algebraic properties of the cross product are the following.

Theorem 2 *If* $u, v, w \in R_F^*$ *then*

(i) $(-u) \odot v = u \odot (-v) = -(u \odot v)$;

(ii) $u \odot v = v \odot u$;

(iii) $(u \odot v) \odot w = u \odot (v \odot w)$;

(iv) *If* u *and* v *have the same sign then* $(u+v) \odot w = (u \odot w) + (v \odot w)$;

(v) $(u \odot v)^{\odot n} = u^{\odot n} \odot v^{\odot n}, \forall n \in N^*$ *, where* $a^{\odot n} = \underbrace{a \odot ... \odot a}_{n \text{ times}}$ *for any* $a \in R_F^*$.

Remark 1) If u is positive and v negative (or u is negative and v positive) then the property of distributivity in (iv) is not verified even if u and v are real numbers.

2) The above properties (i)-(iii) hold for the usual product " $\cdot$ " based on the extension principle. The property (iv) holds in a weaker form: If u and v are on the same side of 0 then for any $w, w \prec 0$ or $0 \prec w$ we have $(u+v) \cdot w = (u \cdot w) + (v \cdot w)$.

The so-called L - R fuzzy numbers are considered important in fuzzy arithmetic. These and their particular cases triangular and trapezoidal fuzzy numbers are used almost exclusively in applications.

Definition 5 ([5], p. 54, [14]) Let $L, R : [0,+\infty) \to [0,1]$ be two continuous, decreasing functions fulfilling $L(0) = R(0) = 1, L(1) = R(1) = 0$, invertible on $[0,1]$. Moreover, let a^1 be any real number and suppose $\underline{a}, \overline{a}$ be positive numbers. The fuzzy set $u : \mathbf{R} \to [0,1]$ is an L - R fuzzy number if

$$u(t) = \begin{cases} L\left(\frac{a^1 - t}{\underline{a}}\right), & \text{for } t \le a^1 \\ R\left(\frac{t - a^1}{\overline{a}}\right), & \text{for } t > a^1. \end{cases}$$

Symbolically, we write $u = \left(a^1, \underline{a}, \overline{a}\right)_{L,R}$, where a^1 is called the mean value of u, $\underline{a}, \overline{a}$ are called the left and the right spread. If u is an L - R fuzzy number then (see e. g. [13])

$$[u]^r = \left[a^1 - L^{-1}(r)\underline{a}, a^1 + R^{-1}(r)\overline{a}\right].$$

Theorem 3 *If u and v are strict positive L - R fuzzy numbers then $u \odot v$ is a strict positive L - R fuzzy number.*

Since we are interested mainly in the applications of the cross product we may restrict our attention to positive fuzzy numbers, however in other cases some similar properties can be obtained (see [3]).

The cross product verifies the following metric property.

Theorem 4 *If u, v have the same sign and $w \in R_F^*$ then*

$$D(w \odot u, w \odot v) \le K_w D(u,v),$$

where $K_w = \max\{\left|\overline{w}^1\right|, \left|\underline{w}^1\right|\} + \overline{w}^0 - \underline{w}^0.$

By using the previous metric property, several properties can be obtained with respect to continuity, differentiability (using the H-differential) and integrability of the product of fuzzy-number-valued functions (see [2]).

The following interpretation related to error theory is a further theoretical motivation of the use of the cross product of fuzzy numbers. Indeed, the consistency of the cross product with the classical theory motivates its use in the case of modelling uncertain data (uncertainty being due to errors of measurement).

We introduce two kinds of errors of fuzzy numbers corresponding to absolute error and relative error in classical error theory and we study these with respect to sum and cross product.

Definition 6 Let u be a fuzzy number. The crisp number $\Delta_L^r(u) = \underline{u}^1 - \underline{u}^r$ is called r - error to left of u and the crisp number $\Delta_R^r(u) = \overline{u}^r - \overline{u}^1$ is called r - error to right of u, where $r \in [0,1]$. The sum $\Delta^r(u) = \Delta_L^r(u) + \Delta_R^r(u)$ is called r -error of u.

If u expresses the fuzzy concept A then $\Delta_L^r(u)$ and $\Delta_R^r(u)$ can be interpreted as the values of tolerance of level r from the concept A to left and to right, respectively. For example, if the triangular fuzzy number $u = (5,7,9)$ expresses "early morning" then $\Delta_L^{\frac{1}{2}}(u) = 1$ (one hour) is the tolerance of level $\frac{1}{2}$ of u towards night from the concept of "early morning" and $\Delta_R^{\frac{1}{4}}(u) = 0.5$ (30 minutes) is the tolerance of level $\frac{1}{4}$ of u towards moon from the concept of "early morning".

A new argument in the use of addition of fuzzy numbers as extension (by Zadeh's principle) of real addition is the validity of the formula

$$\Delta^r(u+v) = \Delta^r(u) + \Delta^r(v)$$

which is consistent to the classical error theory. It is an immediate consequence of the obvious formulas

$$\Delta_L^r(u+v) = \Delta_L^r(u) + \Delta_L^r(v)$$

and

$$\Delta_R^r(u+v) = \Delta_R^r(u) + \Delta_R^r(v).$$

Now, let us study the relative error of the cross product.

Definition 7 Let u be a fuzzy number such that $\underline{u}^1 \neq 0$ and $\overline{u}^1 \neq 0$. The crisp numbers $\delta_L^r(u) = \frac{\Delta_L^r(u)}{|\underline{u}^1|}$ and $\delta_R^r(u) = \frac{\Delta_R^r(u)}{|\overline{u}^1|}$ are called relative r -errors of u to left and to right. The quantity $\delta^r(u) = \delta_L^r(u) + \delta_R^r(u)$ is called relative r -error of u.

Theorem 5 *If u and v are strict positive or strict negative fuzzy numbers then*

$$\delta^r(u \odot v) = \delta^r(u) + \delta^r(v).$$

Corollary 2 *If* u *is a strict positive fuzzy number then* $\delta_L^r\left(u^{\odot n}\right)=n\delta_L^r\left(u\right)$, $\delta_R^r\left(u^{\odot n}\right)=n\delta_R^r\left(u\right)$ *and* $\delta^r\left(u^{\odot n}\right)=n\delta^r\left(u\right)$.

The above theorems show us that the cross product is consistent with the classical error theory (the propagation of errors is governed by a similar law as in the classical case).

4 New Trapezoidal Approximation of the Product

As we have discussed in the previous sections, the usual (Zadeh's extension principle-based) product do not preserve the shape of the operands. Thus, the result of the product, for computational purposes has to be approximated by a trapezoidal number. A new, axiomatic approach to this problem has been introduced in [9]. We will call the trapezoidal approximation of the product based on this method as new trapezoidal approximation of the product.

The method proposed in [9] gives the best approximation of the product under some appropiate conditions. These conditions are natural, so in the approximation of the product the result obtained is motivated from the theoretical point of view. Let us regard the trapezoidal approximation as an operator T, $T : R_F \rightarrow R_F$, which for a given fuzzy number u gives its trapezoidal approximation. The list of requirements which have to be satisfied by an operator of this type are given in [9] below.

1 Conserving some fixed α -cut. E.g. if the operator preserves the 0 and 1 level sets, then the old trapezoidal approximation is reobtained.

2 Invariance to translation of the operator T.

3 Invariance with respect to rescaling.

4 Monotonicity with respect to inclusion.

5 Idempotency (i.e. the trapezoidal approximation of a trapezoidal number is itself).

6 To be best approximation, that is, it should be the nearest in some prescribed sence ($D(T(u),u)\leq D(x,u)$ for any trapezoidal number x).

7 Conserves the so-called expected interval, that is the original fuzzy number and its approximation have the same expected interval (Let us recall here that the expected interval of $u\in R_F$ is $\left[\int_0^1\underline{u}^r dr,\int_0^1\overline{u}^r dr\right]$.

8 Continuity.

9 Compatiblity with the extension principle.

10 Monotonicity with respect to some ordering between fzzy numbers.

11 Invariance with respect to correlation (see [7]).

In [9] the authors propose a trapezoidal approximation which is best approximation and preserves the expected interval, that is conditions 6 and 7 are required. In this case, for $u \in R_F$ we obtain the trapezodal fuzzy number (t_1, t_2, t_3, t_4), where

$$\begin{aligned} t_1 &= -6\int_0^1 r\underline{u}^r dr + 4\int_0^1 \underline{u}^r dr, \\ t_2 &= 6\int_0^1 r\underline{u}^r dr - 2\int_0^1 \underline{u}^r dr, \\ t_3 &= 6\int_0^1 r\overline{u}^r dr - 2\int_0^1 \overline{u}^r dr, \\ t_4 &= -6\int_0^1 r\overline{u}^r dr + 4\int_0^1 \overline{u}^r dr. \end{aligned}$$

In [9], the authors proved also that conditions 2,3,4,5,8,9,10,11 are fulfiled. Let us remark also, that the expected value of a fuzzy number in the sense of [hei92] (this is called in several works defuzzification by center of area method) is the same as the expected value of its trapezoidal approximation. These results lead us to the conclusion that the approximation method proposed in [9] can be useful for applications in geology.

5 Applications of the Product-type Operations in Geology

Recently, fuzzy arithmetic has found several applications in geology (see [1]). In the above cited work the usual (Zadeh's extension principle based) product is used in applications concerning nuclear safety assessment of a repository and for estimation of resources of solid mineral deposits. In this section we propose an alternative study of the same problems, by using the cross product and the new trapezoidal approximation.

The reasons of the possible usefulness of the cross product are the followings. Firstly, in this case the shape of the result of the product is conserved, i.e. the product of triangular numbers is triangular and the product of trapezoidal numbers is trapezoidal. Secondly, the 1-level sets are better taken into account by the use of cross product. Also, the consistency of the cross product with the classical error theory motivate this study.

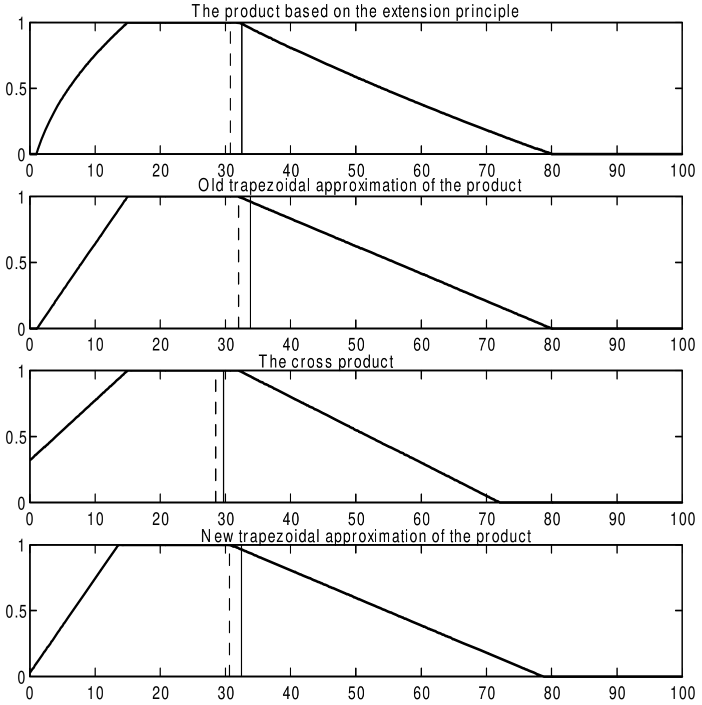


Figure 1

The product of two trapezoidal fuzzy numbers obtained by the different approaches discussed in the paper

The reasons for the possible usefulness of the new trapezoidal approximation are the properties presented in the previous section, that is it preserves the expected interval (also the expected value) and it is the best approximation in some sense.

The product-type operations between fuzzy numbers appear also in several new research fields of geology, such as safety assessment of spent nuclear fuel repositories. The fuzzy modell of the repository is given by a system of linear, coupled fuzzy differential equations. The initial values of several variables and also several coefficients are subject of non-statistical uncertainty, i.e. these are fuzzy numbers. The product-type operations appear in this case in the equations in several places. Firstly, some coefficients in the equations are multiplied by a fuzzy valued function. Also, in the solution of the system obtained by using the extension principle, several times we have to compute powers and even exponentials of a fuzzy number. The three above discussed possible ways to perform multiplication of two fuzzy numbers are illustrated in Figure 1 for the product of the trapezoidal fuzzy numbers $(1,5,8,10)$, $(1,3,4,8)$. The solid vertical line and the dashed vertical line represent the defuzzified values of the results by centroid and expected values (center of area) methods respectively.

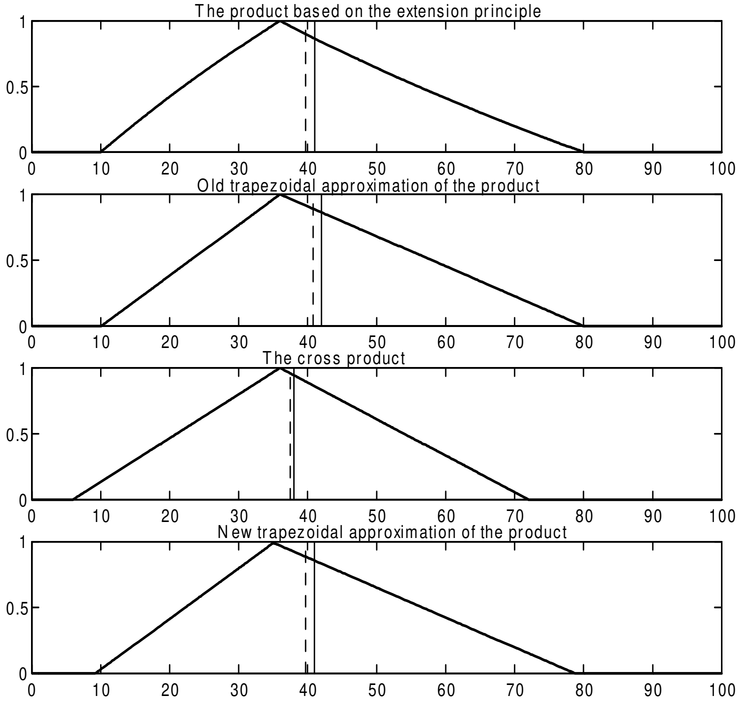


Figure 2
The product of two triangular numbers

In Figure 2, the results of the product of two triangular numbers, $(2,6,8)$, $(5,6,10)$ are presented.

The Figures 1 and 2 do not show a striking difference between the results of the different methods. The difference can be significant if we perform iterative computations with the fuzzy numbers.

In order to show this we consider the exponential functions obtained as power series with respect to the product type operations discussed above. In Figure 3 the results of the exponential-type functions are presented. The fuzzy number considered in the exponent is $(2.2, 4.6, 4.7, 5)$. Solid thin line represent again defuzzification by centroid method, while dashed line the expected value.

Significant difference can be observed between the different results in this case, that is indeed, the iterative use of the product operations leads to different result even after defuzzification. This problem can be avoided by considering and examining all the operations in all the practical problems considered and taking them into account in risk analysis.

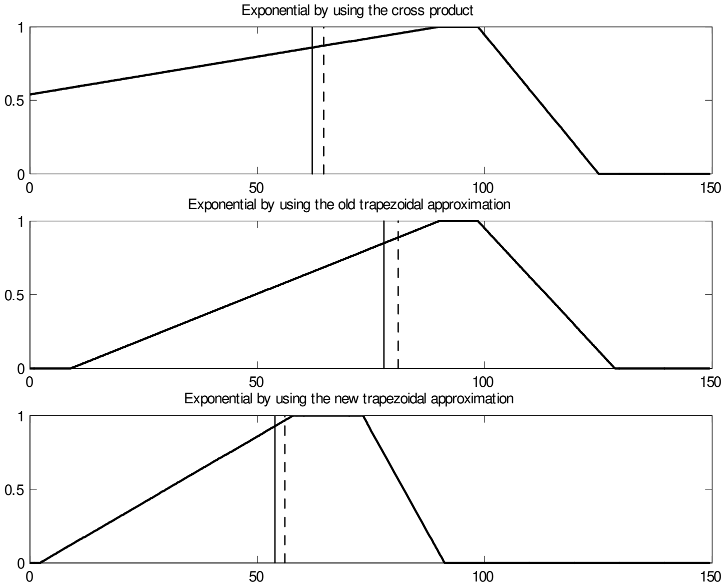


Figure 3

The exponential of a fuzzy number

This figure suggests us also, to be careful with the use of the cross product in the construction of an exponential, since in some cases the result given by this operation is negative, fact which is impossible from a possibilistic point of view. The figure suggests that probably the best behaviour is that of the new trapezoidal approxmation. Unfortunately, since its very high computational complexity this method cannot be effectively used for modeling purposes in nuclear safety assessment.

The next application presents as in [1], resource estimation on several bauxite deposites in Hungary. In the same way as with the traditional methods, the tonnage of the resources is obtained by the product of the deposit area, the average thickness and the average bulk-density of the studied ore or mineral commodity (see Figure 4 and see also [1]). Large deposits can be split into blocks, preferably along natural boundaries, such as tectonic lines.

We present the results obtained by the old trapezoidal approximation, the results obtained by using the cross product and the new trapezoidal approximation compared with the product based on the extension principle (see Figures 5, 6).

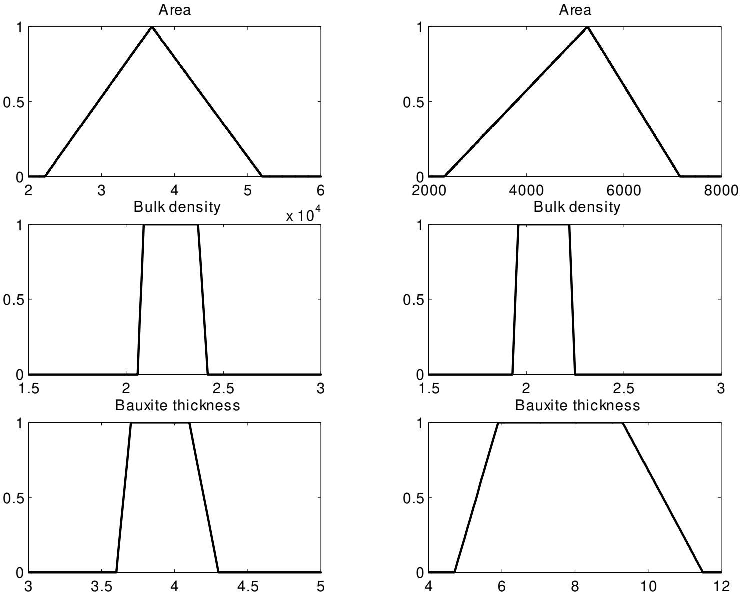


Figure 4
Fuzzy numbers used for the resource calculations at Sőcz-Szárhegy I and I/A (left) and Óbarok IX (right) sites

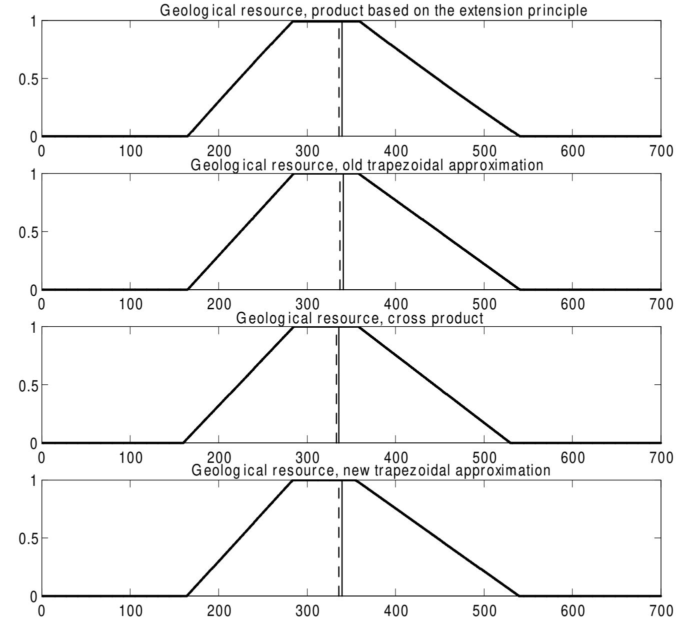
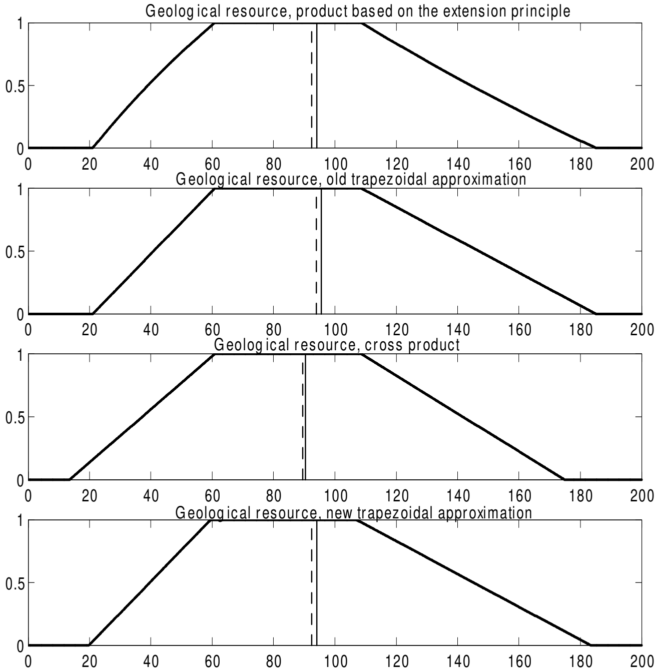


Figure 5
Fuzzy numbers of the tonnage calculation, Szőc-Szárhegy I. and I/A. (thousand tones)

We observe that, if we defuzzify the results obtained by the three different product-type operations we conclude that the results are different. Also, we observe that after defuzzification (by centroid method) the result of the cross product in the study of the deposites at both sites is smaller than that of the new trapezoidal approximation, which is smaller at its turn than the old trapezoidal approximation of the product. So, the risks of an investment at this site can be more realistically evaluated if we use the cross product or the new trapezoidal approximation, however the results are not very different. From the risk analysis point of view of the investment, the most important information is not in the defuzzified value, but in the trapezoidal fuzzy numbers themselves. Indeed, from a possibilistic point of view, the fuzzy numbers contain much more information than only the defuzzified vales and the risks can be better evaluated considering the tonnage as a fuzzy number.


Geological resource, product based on the extension principle

Geological resource, old trapezoidal approximation

Geological resource, cross product

Geological resource, new trapezoidal approximation

Figure 6

Fuzzy numbers of the tonnage calculation, Óbarok IX. (thousand tones)

Conclusions and Further Research

Several methods for the multiplication of fuzzy numbers are discussed from the theoretical and practical point of view. As a conclusion of this research we can state that the theretical properties of the cross-product and the new trapezoidal approximation motivate the usefulness of both methods in geology, however, usually the most conservative method is the old trapezoidal approximation. So, there exist reasons for using all the above mentioned approaches and to take into account the results of all the approaches in a risk analisys or a safety assessment.

From the computational point of view, let us remark that the old trapezoidal approximation and the cross product can be computed in an easy way taking into account only the endpoints of the core and support of the trapezoidal operands and the extension principle can be avoided in these cases. In the case of the new trapezoidal approximation, since the implementation of this operation involves the use of the extension principle and then numerical integration on the side-functions of the fuzzy numbers, the computational complexity is high. This makes almost impossible to use the new trapezoidal approximation in iterative computations.

For further research we propose effective implementation of the new trapezoidal approximation, and the design of new, computationally simple methods for the approximation of the product based on the extension principle.

References

[1] Gy. Bárdossy, J. Fodor, *Evaluation of Uncertainties and Risks in Geology*, Springer Verlag, Berlin Heidelberg, New York, 2004

[2] A. I. Ban, B. Bede, Cross product of fuzzy numbers and properties, *Journal of Fuzzy Mathematics*, to appear

[3] A. I. Ban, B. Bede, Cross product of $L-R$ fuzzy numbers and properties, *Anals of Oradea Univ., Fasc. Matem.* 9(2003), 95-108

[4] M. Delgado, M. A. Vila and W. Voxman, On a canonical representation of fuzzy numbers, *Fuzzy Sets and Systems*, 93(1998), 125-135

[5] D. Dubois and H. Prade, *Fuzzy Sets and Systems*, Academic Press, New York, 1980

[6] R. Fuller and T. Keresztfalvi, On generalization of Nguyen's theorem, *Fuzzy Sets and Systems*, 41(1991), 371-374

[7] R. Fullér, P. Majlender, On interactive fuzzy numbers, *Fuzzy Sets and Systems* 143(2004) 355-369

[8] R. Goetschel and W. Voxman, Elementary fuzzy calculus, *Fuzzy Sets and Systems*, 18(1986), 31-43

[9] P. Grzegorzewski, E. Mrówka, Trapezoidal approximations of fuzzy numbers, *Fuzzy Sts and Systems*, 153(2005), 115-135

[10] S. Heilpern, The expected value of a fuzzy number, *Fuzzy Sets and Systems*, 47(1992), 81-86

[11] Ming Ma, M. Friedman and A. Kandel, A new fuzzy arithmetic, *Fuzzy Sets and Systems*, 108(1999), 83-90

[12] M. L. Puri and D. A. Ralescu, Differentials of fuzzy functions, *J. Math. Anal. Appl.*, 91(1983), 552-558

[13] M. Wagenknecht, R. Hampel and V. Schneider, Computational aspects of fuzzy arithmetics based on Archimedean t-norms, *Fuzzy Sets and Systems*, 123(2001), 49-62

[14] Congxin Wu and Zengtai Gong, On Henstock integral of fuzzy number valued functions (I), *Fuzzy Sets and Systems*, 120(2001), 523-532

[15] L. A. Zadeh, Fuzzy sets, *Inform. and Control*, 8(1965), 338-353

The Pseudooperators in Second Order Control Problems

Márta Takács

Institute of Intelligent Engineering Systems, John von Neumann Faculty of Informatics, Budapest Tech
Bécsi út 96/B, H-1034 Budapest, Hungary
takacs.marta@nik.bmf.hu

Abstract: This paper deals with the control of a dynamic system where the gains of the conventional PD controller are previously chosen by fuzzy methods in such a way as to obtain the optimal trajectory tracking. The gain factors are determined by solving fuzzy equations, and based on the sufficient possibility measure of the solution. It will be shown, that the rule premise for the given system input in fuzzy control system may also determine the possibility of realizing a rule. This possibility can be used for verifying the rule and for changing the rule-output, too. This leads to the optimization of the output. When calculating the possibility value the possible functional relation between the rule-premise and rule-consequence is taken into account. For defining the rule of inference in Fuzzy Logic Control (FLC) system special class of t-norm is used. The proposed fuzzy logic controller uses the functional relation between the rule premises and consequences, and the special class of pseudo-operators in the compositional rule of inference.

Keywords: FLC, fuzzification of the linear equations, PD controllers, pseudo-operators

1 Introduction

There is a question that arises during the studying of fuzzified functions: what are those practical problems where given beside certain fuzzified function parameters, an approximation can be provided to other unknown but also fuzzy-type function parameters. If a scruple, linear function relationship is observed, the fuzzification problem of the

$$K_p e + K_d \dot{e} = y \tag{1}$$

type law of the PD-type controller emerges.

The conventional linguistic FLC uses fuzzified quantities $e, \dot{e}$ (error and error change) as inputs and y as output. The rules of this system are

$$\text{if} \quad e \text{ is E} \quad \text{and} \quad \dot{e} \text{ is } \dot{\text{E}} \quad \text{then} \quad y \text{ is Y}$$

where $\text{E}, \dot{\text{E}}$ and Y are linguistic terms, whose can be NB(negative big), NM(negative medium), NS(negative small), ZE(zero), PS(positive small), PM(positive medium), PB(positive big). Fuzzy membership functions cover linguistic terms. The scaling and normalization of parameters domains are made by experts.

The input variables of a dynamic system to be controlled can be the error (*e*), which is the difference between the desired and the actual output of the system and the errorchange ($\dot{e}$). In a typical PD controller using these variables the *y* output is determined by the control law given by the equation (1), where the control gains K_p,K_d also could be modified during the operation in order to bring the system to be controlled into a desired state. There types of FLCs are called tuning-type controllers. In the literature there are indications regarding the solution of this problem. Some soft computing based techniques have been published for the on-line determination of these gains[1].

In further explanation a possible way for tuning these parameters is given, to achieve an efficient system-performance. The architecture of the proposed controller can be seen in Fig. 1. The conventional PD controller and the Fuzzy Logic Controller (FLC) use the same $e,\dot{e}$ input variables and the FLC also uses the output *y* of the PD controller (this is required because of the linear relationship in $K_p e + K_d \dot{e} = y$). The FLC gives two crisp outputs, the gains K_p ,K_d, to the PD controller that calculates the new *y* by using these gains and $e,\dot{e}$ as inputs. The rules of the FLC are given in the the form of:

$$\text{if} \ \left(e \text{ is E} \quad \text{and} \quad \dot{e} \text{ is } \dot{\text{E}} \quad \text{and} \quad y \text{ is Y}\right) \text{then} \ \left(K_p \text{ is K}_\text{p} \quad \text{and} \quad K_d \text{ is K}_\text{d}\right) \qquad (2)$$

where $\text{E}, \dot{\text{E}}, \text{Y}, \text{K}_\text{p}, \text{K}_\text{d}$ are linguistic terms, which can be for example N(negative), Z(zero), P(positive). Fuzzy membership functions cover linguistic terms. The scaling and normalization of parameters domains are made by experts.

The performance of the propose self tuning controller has been evaluated. For this purpose a second order differential equation has been chosen. The results serves to show the effects of the operators used in the rules of inference as well as the effects of the generalized t-norms.

The theoretical background of the membership functions of the linguistic terms is given, and the applied generalized t-norms and their generator functions are summarized, based on the general theoretical publication [2], [3]. A theoretical interpretation of possibility measure of the rule realization is given using the same generator function as by definition of the membership functions and applied t-norm. The functional dependence used to determine the possibility of the rule plays a very important role. It can be used for the rule base construction, and for the inference mechanism as well by narrowing the linguistic rule consequence. With these narrowing rule consequences a modified FLC model can be constructed, where the rule consequences are surfaces above the K_p,K_d plane.

In the paper a method which using the functional relationship between $K_p, K_d, e, \dot{e}, y$ parameters is presented, that creates the rule base on one hand, and furthermore uses this functional dependence in the inference mechanism too.

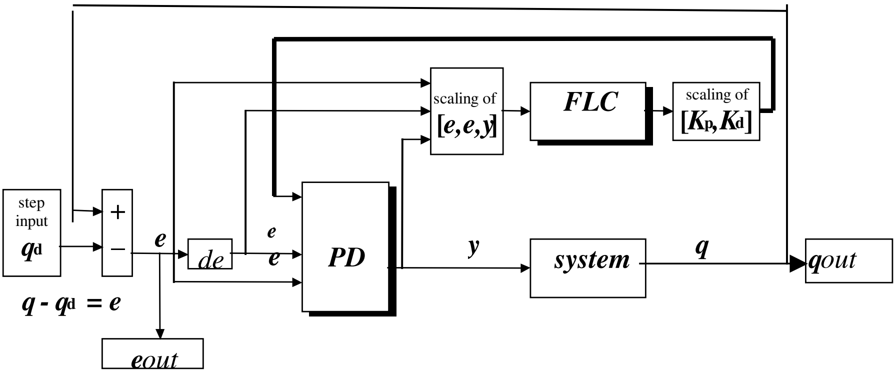


Figure 1

The architecture of the proposed controller

Figure 1 illustrates how such a tuning FLC can be integrated into the system. The conventional PD controller and the FLC has the same $e, \dot{e}$ as inputs and the tuning FLC also uses the output *y* of the conventional controller (this is required because of the linear relationship in (1)). The FLC gives two crisp outputs to the PD controller that calculates a new *y* by using these gains and $e, \dot{e}$ as inputs.

Following the procedure of FLC construction, contains of steps:

st1 determination of fuzzification strategy

st2 the choise of quantities to be fuzzified

st3 fuzzification of these quantities and the rule base construction

st4 choosing of inference mechanism

st5 choosing of defuzzification model,

these general steps cover different mathematical procedures depending on the choice of strategy. This paper presents two procedures on an example: a Mamdani-type, in which a novel construction of the rule-base is given, and another one which is said to be *possibility-modified* and the rule possibilities integrated into the rule outputs.

2 General Concept

2.1 Special Types of the Fuzzy Numbers

A fuzzy subset *A* of a universe of discourse *X* is defined as $A = \{(x, \mu(x)) | x \in X, \mu_A : X \to [0,1]\}$. Denote *FX* the set of all fuzzy subsets of *X*. The characteristic function of *A* will be denoted by χ_A. If the universe is $X = \Re$, and we have a membership function

$$A(x) = \begin{cases} g^{(-1)}\left(\dfrac{|x-\alpha|}{\delta}\right), & \text{if } \delta \neq 0 \\ \chi_\alpha(x), & \text{if } \delta = 0 \end{cases} \tag{3}$$

$\alpha \in \Re, \delta \geq 0$, then the fuzzy set given by $A(x)$ will be called quasitriangular fuzzy number with the center α and width δ, and we will recall for it by QTFN(α,δ).

2.2 Pseudo-operators

Generally details about pseudo-analysis and pseudo-operators we can read in [4], [5]and [6].

Pseudo-analysis

The base for the pseudo-analysis is a real semiring, defined in the following way:

Let $[a,b]$ be a closed subinterval of $[-\infty,+\infty]$ (in some cases semi-closed subintervals will be considered) and let $\preceq$ be a total order on $[a,b]$. A *semiring* is the structure $(\preceq, \oplus, \otimes)$ if the following hold:

- $\oplus$ is pseudo-addition, i.e., a function $\oplus : [a,b] \times [a,b] \to [a,b]$ which is commutative, non-decreasing (with respect to $\preceq$), associative and with a zero element denoted by $\mathbf{0}$;
- $\otimes$ is pseudo-multiplication, i.e., a function $\otimes : [a,b] \times [a,b] \to [a,b]$ which is commutative, positively non-decreasing ($x \preceq y$ implies $x \otimes z \preceq x \otimes y$ where $z \in [a,b]_+ = \{z | z \in [a,b], \mathbf{0} \preceq z\}$ associative and for which there exists a unit element denoted by $\mathbf{1}$.
- $\mathbf{0} \otimes z = \mathbf{0}$
- $x \otimes (y \oplus z) = (x \otimes y) \oplus (x \otimes z)$

Three basic classes of semirings with continuous (up to some points) pseudo-operations are:

(i) The pseudo-addition is an idempotent operation and the pseudo-multiplication is not.

(ii) Semi-rings with strict pseudo-operations defined by a monotone and continuous generator function $g : [a,b] \to [0,+\infty]$, i.e., *g*-semirings:

$$x \oplus y = g^{-1}(g(x) + g(y)) \text{ and } x \otimes y = g^{-1}(g(x)g(y)).$$

(iii) Both operations, $\oplus$ and $\otimes$, are idempotent.

More on this structure can be found in [4].

T-norm as the Pseudooperator

Let *T*: $\mathrm{I}^2 \to \mathrm{I}$, (I=[0,1]) be a a t-norm. The t-norm is Archimedean if and only if it admits the representation $T(a,b) = g^{-1}(g(a) + g(b))$, where the generator function $g{:}\,\mathrm{I} \to \Re^+$ is continuous, strictly decreasing function, with the boundary conditions , $g(0) = 1, g(1) = 0$ and let

$$g^{(-1)}(x) = \begin{cases} g^{-1}(x) & x \in \mathrm{I} \\ 0 & x \notin \mathrm{I} \end{cases} \tag{4}$$

the pseudoinverse of the function *g*. The generalization of this representation is

$$T_{gp}(a,b) = g^{-1}\left(g^p(a) + g^p(b)\right)^{\frac{1}{p}}, \tag{5}$$

and it can be said, that the T_{gp} function is an Archimedean t-norm given by generator function g^p , $p \in [1,\infty)$.

t-norms were introduced as binary operations. Since they are associative, they also can be considered as operations with more than two arguments.

(i) The associativity (T2) allows us to extend each *t*-norm *T* in a unique way to an *n*-ary operation by induction, defined for each *n*-tuple $(x_1, x_2, ...x_n) \in [0,1]^n$, $(n \in N \cup \{0\})$as

$$\mathop{T}_{i=1}^{0} x_i = 1, \ \mathop{T}_{i=1}^{n} x_i = T\left(\mathop{T}_{i=1}^{n-1} x_i, x_n\right) = T(x_1, x_2, ...x_n)$$

If, in a specific case, we have $x_1 = x_2 = ... = x_n = x$, in short, it can be written in the form $x_T^{(n)}$ instead of $T(\underbrace{x, x, ...x}_{n})$.

(ii) The fact that each *t*-norm T is weaker than the minimum T_M makes it possible to extend it to a countable infinity operation, putting for each $(x_i)_{i\in N} \in [0,1]^N$

$$\mathop{T}_{i=1}^{\infty} x_i = \lim_{n\to\infty} \mathop{T}_{i=1}^{n} x_i$$

Note that the limit on the right side always exists, since the sequence

$$\left(\mathop{T}_{i=1}^{n} x_i\right)_{n\in N}$$

is non-increasing and bounded from below.

Continuity of the T-norms

In general, a real function of two variables, e.g., with the domain $[0,1]^2$ may be continuous in each variable without being continuous on $[0,1]^2$. Triangular norms and conorms are exceptions from this.

Proposition. A t-norm is continuous if and only if it is continuous in its first component, i.e., for each fixed $y \in [0,1]$ the one-place function $T(\cdot, y) : [0,1] \to [0,1]$, or briefly $x \to T(x, y)$ is continuous.

Keeping in mind that the only extra property for the t-norm T was the monotonicity, the proposition can be extended for any *monotone* function F of two variables, which is continuous in both components.

For applications quite often *weaker forms of continuity* are sufficient.

2.3 FLC Systems

One rule in a FLC system has form: if $\text{x is } A(x)$ then $\text{y is } B(y)$, where x is the system input, y is the system output, $\text{x is } A(x)$ is the rule-premise, $\text{y is } B(y)$ is the rule-consequence. *A*(*x*) and *B*(*y*) are linguistic terms and they can be described by QTFN-s.

For a given input fuzzy set $A'(x)$, in a mathematical-logical sense, the output fuzzy set $B'(y)$, the model of the compositional rule of inference in the Mamdani type controller will be generated with a Generalized Modus Ponens (GMP):

$$B'(y)=\sup_{x\in X} T\big(T(A(x),A'(x)),B(y)\big)=T\left(\underbrace{\sup_{x\in X} T(A(x),A'(x)),}_{DOF}B(y)\right) \tag{6}$$

where DOF is the degree of firing value for the rule.

2.4 The Fuzzification of the Linear Equalities

Possibility Logic

In possibility logic the propositions can be true or false, but we do not know exactly their truth value. If we know the truth value of several of them, then we can infer the truth value of more complex terms.

Every proposition p∈P, (where the (P,∨,∧,¬) is a Boolean-algebra) has:

- Possibilistic measure: Poss(p)
- Necessity measure: Nec(p), with the following properties:
- Poss (false)=0=Nec(false)
- Poss(true)=1=Nec(true)

  ∀p,q∈P , Poss(p∨q)=max(Poss(p),Poss(q))

  ∀p,q∈P , Nec(p∧q)=min(Nec(p),Nec(q))

  ∀p∈P , Nec(p)=1-Poss(¬p)

The essential consequences are:

max(Poss(p),Poss(¬p))=1

min(Nec(p),Nec(¬p))=0

Nec(p)≤Poss(p).

For us the following properties are important

Poss(p∧q) ≤ min(Poss(p),Poss(q))

Nec (p∨q) ≥ max(Nec(p),Nec(q))

For $\alpha,\beta\in$I

$$\alpha \operatorname{Im} \beta = \begin{cases} 0 & \text{if } \alpha+\beta \le 1 \\ \beta & \text{if } \alpha+\beta > 1 \end{cases}$$

is a monotone increasing operation. It is easy to see, that

Poss(p∧q) ≥ Nec(p) Im Poss(q).

Possibility Measure of the Control Law

The *e,* $\dot{e}$*,y* values in (1) are uncertain, so they can be replaced by the quasitriangular fuzzy numbers,

$$\widehat{e}(e)=g^{(-1)}\left(\frac{|e-e_v|}{\delta_1}\right),\ \widehat{\mathrm{e}}(\mathrm{e})\ \text{is a}\ (e_v,\delta_1)\ \text{type fuzzy number}$$

$$\widehat{\dot{\mathrm{e}}}(\dot{\mathrm{e}})=g^{(-1)}\left(\frac{|\dot{e}-\dot{e}_v|}{\delta_2}\right),\ \widehat{\dot{\mathrm{e}}}(\dot{\mathrm{e}})\ \text{is a}\ (\dot{e}_v,\delta_2)\ \text{type fuzzy number}$$

$$\widehat{y}(z)=g^{(-1)}\left(\frac{|z-y_v|}{\delta_3}\right),$$ $\widehat{y}$ (z) is a (y_v,δ_3) type fuzzy number, (δ_1,δ_2 ,δ_3>0).

If δ_1,δ_2 ,δ_3=0, then $\widehat{e}(e)=\chi_{e_v}(e),\widehat{\dot{e}}(\dot{e})=\chi_{e_v}(\dot{e})$, $\widehat{y}(z)=\chi_{y_v}(z)$ are singletons, given by characteristic functions.

$$K_p e+K_d\dot{e}-y\cdot 1=0$$

We can give a (g,p,δ) fuzzification of this equality:

$$\sigma(K_p,K_d)=g^{(-1)}\left(\frac{|K_p e_v-K_d\dot{e}_v-y_v|}{\|(K_p\delta_1,K_d\delta_2,\delta_3)\|_q}\right)$$

σ(K_p,K_d) is a possibilistic measure of equality

$$E\widehat{Q}\left(\widehat{l}\left(\left(\widehat{e},\widehat{\dot{e}},\widehat{y}\right),\left(K_p,K_d\right)\right),\chi_0\right)=\sigma\left(K_p,K_d\right),\ \text{i.e.}$$

$$Poss(K_p\widehat{e}+K_d\widehat{\dot{e}}=\widehat{y})=\sigma\left(K_p,K_d\right).$$

For fixed K_p,K_d values we have an interpreted term, and σ(K_p,K_d) is the truth value for them in possibility logic. K_p,K_d have to be choosen to provide the maximum value of σ(K_p,K_d) which is equal to 1.

A conventional IF ... THEN rule to determine K_p,K_d in case of given *e,* $\dot{e}$*,y* is the following:

$$IF\ \widehat{e}(e)\,AND\,\widehat{\dot{e}}(\dot{e})\,AND\ \widehat{y}(z)THEN\ \widehat{K}_p\left(K_p\right)AND\ \widehat{K}_d\left(K_d\right).$$

The possibility of realization of this rule by using these inputs and taking into account the linear relatian between the parametrs (Eq. (1)) results in the modified rule

$$IF\ \widehat{e}(e)\,AND\,\widehat{\dot{e}}(\dot{e})\,AND\ \widehat{y}(z)THEN\ \sigma\left(\widehat{K}_p\left(K_p\right),\widehat{K}_d\left(K_d\right)\right).$$

Possibility and Necessity Measures of the Equation System

If some dynamically not coupled systems, given in Fig. 2 work simultanously then an equation type (1) can be related to each system. The possibility measures of realization of an equation by fixed K_p, K_d is

$$Poss(K_p\hat{e} + K_d\hat{\dot{e}} = \hat{y}) = \sigma\left(K_p, K_d\right).$$

If we have two equations, we have $\sigma_1(K_{1p},K_{1d})$, $\sigma_2(K_{2p},K_{2d})$, and the possibility measure of the simultaneous realization of these equations and simultaneous working of systems by the actual parameter values is

$$\sigma((K_{1p},K_{1d}) \wedge (K_{2p},K_{2d})) \leq \min(\sigma_1(K_{1p},K_{1d}),\ \sigma_2(K_{2p},K_{2d}))$$

and furthermore

$$\sigma_{1(}(K_{1p},K_{1d}) \wedge (K_{2p},K_{2d})) \geq \mathrm{Nec}(K_{1p},K_{1d}) \text{ Im } \sigma_2(K_{2p},K_{2d})=$$

$$\begin{cases} 0 & \text{if } \mathrm{Nec}\left(\mathbf{K_{1p}},\mathbf{K_{1d}}\right)+\sigma\left(\mathbf{K_{2p}},\mathbf{K_{2d}}\right)\leq 1 \\ \sigma\left(\mathbf{K_{2p}},\mathbf{K_{2d}}\right) & \text{if } \mathrm{Nec}\left(\mathbf{K_{1p}},\mathbf{K_{1d}}\right)+\sigma\left(\mathbf{K_{2p}},\mathbf{K_{2d}}\right)> 1 \end{cases} =$$

$$\begin{cases} 0 & \text{if } \sigma\left(\mathbf{K_{2p}},\mathbf{K_{2d}}\right)\leq \sigma\left(\neg\left(\mathbf{K_{1p}},\mathbf{K_{1d}}\right)\right) \\ \sigma\left(\mathbf{K_{2p}},\mathbf{K_{2d}}\right) & \text{if } \sigma\left(\mathbf{K_{2p}},\mathbf{K_{2d}}\right)> \sigma\left(\neg\left(\mathbf{K_{1p}},\mathbf{K_{1d}}\right)\right) \end{cases}$$

$$\sigma((K_{2p},K_{2d}) \wedge (K_{1p},K_{1d})) \geq \mathrm{Nec}\ (K_{2p},K_{2d}) \text{ Im } \sigma_2(K_{1p},K_{1d}) =$$

$$= \begin{cases} 0 & \text{if } \sigma\left(\mathbf{K_{1p}},\mathbf{K_{1d}}\right)\leq \sigma\left(\neg\left(\mathbf{K_{2p}},\mathbf{K_{2d}}\right)\right) \\ \sigma\left(\mathbf{K_{1p}},\mathbf{K_{1d}}\right) & \text{if } \sigma\left(\mathbf{K_{1p}},\mathbf{K_{1d}}\right)> \sigma\left(\neg\left(\mathbf{K_{2p}},\mathbf{K_{2d}}\right)\right) \end{cases}$$

The necessity of the simultaneous realization of these equations is analogously

$$\mathrm{Nec}\ ((K_{1p},K_{1d}) \vee (K_{2p},K_{2d})) \geq \max(\mathrm{Nec}\ (K_{1p},K_{1d})\ ,\ \mathrm{Nec}(K_{2p},K_{2d}))$$

when the Nec (K_{1p},K_{1d}) and Nec(K_{2p},K_{2d}) are the given as heuristic necessity measures by experts.

The Possibility Model of the Problem (1)

The membership function of the t-norm of fuzzy sets is defined as follows

$$\mu(x) \cap \nu(x) = T(\mu(x), \nu(x)) \in \mathcal{FR} \qquad (7)$$

The Mamdani type controller applies the rule: if x is $\mu(x)$ then y is $\nu(y)$, where x is the system input, y is the system output, x is $\mu(x)$ is the rule-premise, y is $\nu(y)$ is the rule-consequence. μ(*x*) and ν(*y*) are linguistic terms and they can be described by QTFN-s.

For a given input fuzzy set $\mu'(x)$, in a mathematical-logical sense, the output fuzzy set $\nu'(y)$, will be generated with a Generalized Modus Ponens (GMP).

At every fixed $x\in\Re$ a T-fuzzification of the function value of the parametric function *f(a$_1$,a$_2$,...a$_k$, x)* by the fuzzy parameter vector $\mu_a=(\mu_1,\mu_2,...\mu_k)$ is a fuzzy set of FR..

The (g,p,δ) fuzzification of a linear equality $\alpha_1x_1+\alpha_2x_2+...+\alpha_nx_n=\alpha_0$ by the fuzzy vector parameter $\mu_a=(\mu_1,\mu_2,...\mu_n)$ (where the coefficients α_i are uncertainly parameters, and replaced by $\mu_i(\alpha_i,\delta_i)$ QTFN-s, and the fuzzification of function will be defined by T_{gp} norm), is

$$\sigma(x)=g^{(-1)}\left(\frac{|l(\alpha,x)|}{\|diag(\delta)\cdot x\|_q}\right)=g^{(-1)}\left(\frac{|l(\alpha_1x_1+\alpha_2x_2+...+\alpha_nx_n-\alpha_0)|}{\|diag(\delta)\cdot x\|_q}\right) \qquad (8)$$

where

$$\|(v_1,v_2,...v_n)\|_p=\begin{cases}\left(\sum\limits_{j=1}^{p}|v_j|^p\right)^{1/p} & 1\le p<\infty\\ \max\limits_{j}|v_j| & p=\infty\end{cases},\qquad q=\begin{cases}1 & \text{if } p=\infty\\ \infty & \text{if } p=1\\ \dfrac{p}{p-1} & \text{otherwise}\end{cases}$$

and diag(δ) is a diagonal matrix from elements δ_i. $\sigma(x)$ will be called *possibility measure* of equality [2]. All the properties, written in the section 2.2. are necessery for the proof of the propositions above.

Let be T_{gp} an Archimedian t-norm given by generator function g^p, $p\in[1,\infty)$.

The membership function of the t-norm of fuzzy sets is defined as follows

$$\mu(x)\cap\nu(x)=T(\mu(x),\nu(x))\in FR \qquad (9)$$

The Mamdani type controller applies the rule: if x is $\mu(x)$ then y is $\nu(y)$, where x is the system input, y is the system output, x is $\mu(x)$ is the rule-premise, y is $\nu(y)$ is the rule-consequence. $\mu(x)$ and $\nu(y)$ are linguistic terms and they can be described by QTFN-s.

For a given input fuzzy set $\mu'(x)$, in a mathematical-logical sense, the output fuzzy set $\nu'(y)$, will be generated with a Generalized Modus Ponens (GMP).

At every fixed $x\in\Re$ a T-fuzzification of the function value of the parametric function *f(a$_1$,a$_2$,...a$_k$, x)* by the fuzzy parameter vector $\mu_a=(\mu_1,\mu_2,...\mu_k)$ is a fuzzy set of FR..

Let EQ be a non-fuzzy equality relation on universe. The T-fuzzification of EQ is a fuzzy set on $\mathcal{FR}\times\mathcal{FR}$

$$\widehat{EQ}(\mu(x),\nu(y))=\sup_{x=y}T(\mu(x),\nu(y))=\sup_{x}T(\mu(x),\nu(x)) \tag{10}$$

The (g,p,δ) fuzzification of a linear function $l(\alpha,x)=\alpha_1x_1+\alpha_2x_2+...+\alpha_nx_n$ by the fuzzy vector parameter $\mu_a=(\mu_1,\mu_2,...\mu_n)$ (where the coefficients α_i are uncertainly parameters, and replaced by QTFN(α_i,δ_i), and the fuzzification of function will be defined by T_{gp} norm), is given in [4].

The (g,p,δ) fuzzification of a linear equality $\alpha_1x_1+\alpha_2x_2+...+\alpha_nx_n=\alpha_0$ by the fuzzy vector parameter $\mu_a=(\mu_1,\mu_2,...\mu_n)$ is:

$$\widehat{EQ}\left(\hat{l}(\mu_a,x),\chi_0\right)=\hat{l}(\mu_a,x)(0)=$$

$$\sigma(x)=g^{(-1)}\left(\frac{|l(\alpha,x)|}{\|diag(\delta)\cdot x\|_q}\right)=g^{(-1)}\left(\frac{|l(\alpha_1x_1+\alpha_2x_2+...+\alpha_nx_n-\alpha_0)|}{\|diag(\delta)\cdot x\|_q}\right) \tag{11}$$

$\sigma(x)$ will be called *possibility measure* of equality [2],[3].

3 Construction of the Mamdani-type FLC for Control Law, Example

st1 Let us chose a Mamdani-type linguistic model for the problem (1).

st2 $e,\dot{e},y$ quantities are uncertain, fuzzified, and comprise the FLC and the rule-inputs. K_p, K_d are also uncertain, fuzzified but they comprise the outputs. The rule type for the scaling of the gain papameters K_p, K_d is

$$\text{if } (e \text{ is E} \quad \text{and} \quad \dot{e} \text{ is } \dot{\text{E}} \quad \text{and} \quad y \text{ is Y}) \quad \text{then} \quad (K_p \text{ is K}_\text{p} \quad \text{and} \quad K_d \text{ is K}_\text{d})$$

shortly

$$\text{if} \quad \text{E}\cap\dot{\text{E}}\cap\text{Y}(\underline{e}) \quad \text{then} \quad \text{K}_\text{p}\cap\text{K}_\text{d}(\underline{k}) \quad \text{or} \quad \text{if} \quad \text{E}\cap\dot{\text{E}}\cap\text{Y}(\underline{e}) \quad \text{then} \quad \text{K}(\underline{k}) \ .$$

(Details see in [4])

st3 Experts can provide those $[-L_e,L_e],[-L_{\dot{e}},L_{\dot{e}}]$ intervals where $e,\dot{e}$ quantities exist. For simplification and generalization of the problem these $[-L_e,L_e],[-L_{\dot{e}},L_{\dot{e}}]$ intervals are normalized and transformed into interval [-1,1]. During the scaling operation $e,\dot{e}$ receive 77 linguistic terms, there being determined by (3) type fuzzy numbers, for example

$$\mathrm{E}(e)=\begin{cases} g^{(-1)}\left(\dfrac{|e-e_c|}{de}\right), & \text{if de} \neq 0 \\ \chi_{e_c}(e), & \text{if de} = 0 \end{cases}$$

$e\backslash\dot{e}$	**NB**	**NM**	**NS**	**ZE**	**PS**	**PM**	**PB**
NB	NB	NB	NB	NM	NM	NS	ZE
NM	NB	NB	NM	NM	NS	ZE	PS
NS	NB	NM	NM	NS	ZE	PS	PM
ZE	NM	NM	NS	ZE	PS	PM	PM
PS	NM	NS	ZE	PS	PM	PM	PB
PM	NS	ZE	PS	PM	PM	PB	PB
PB	ZE	PS	PM	PM	PB	PB	PB

Table 1

These 49 possibilities would increase seven times if the *y* quantity was normalized and scaled likewise. It should be noted, however, that $e,\dot{e},y$ quantities are not independent from each other. The relationship generally used by experts in such controllers, (see *Table 1* for the y quantity), can be applied for completing input parameters into the rule. Finally we have 49 different rule inputs. The scaling of *y* is the same on normalized interval [-1,1].

For the rule outputs also linguistic terms are defined which are obtained within the domain of K_p,K_d by scaling. The $\lfloor -L_{K_p},L_{K_p}\rfloor,\lfloor -L_{K_d},L_{K_d}\rfloor$ intervals and the scaling are determined by experts. For the given $\mathrm{E},\dot{\mathrm{E}},\mathrm{Y}$ the suitable $\mathrm{K_p,K_d}$ rule outputs are chosen based on experience meta-rules or tiresome experimental work.

In our case the $\mathrm{K_p,K_d}$ output fuzzy domains will be determined as such for which the possibility of law (1) is the greatest, in case of given $\mathrm{E},\dot{\mathrm{E}},\mathrm{Y}$.

First let us assign linguistic terms to $\mathrm{K_p,K_d}$ (like by $e,\dot{e},y$) on $\lfloor -L_{K_p},L_{K_p}\rfloor\lfloor -L_{K_d},L_{K_d}\rfloor$ interval. The possible K_p is $\mathrm{K_p}$ and K_d is $\mathrm{K_d}$ (i.e. $\mathrm{K_p}\cap\mathrm{K_d}$) domain-number is 49.

Define the possibility measure:

$$\sigma\left(K_p,K_d\right)=g^{(-1)}\left(\frac{\left|K_p e_c+K_d\dot{e}_c-y_c\right|}{\left\|diag(\delta)\cdot\left[K_p,K_d,1\right]^T\right\|_q}\right) \tag{12}$$

for each rule-premise . The possibilistic rule is defined as follows:

if $\mathrm{E}\cap\dot{\mathrm{E}}\cap\mathrm{Y}(\underline{e})$ then $\sigma\left(K_p,K_d\right)$ or if $\mathrm{E}\cap\dot{\mathrm{E}}\cap\mathrm{Y}(\underline{e})$ then $\sigma(\underline{k})$

In principle, any $K_p \cap K_d$ intersection can be assigned as output to the rule-premise, but in our case the one with the greatest possibility is used, i.e.

$$poss(\,i\,max, j\,max) = \max_{i,j}\left(\min_{\underline{k}}\left(\sigma(\underline{k}) \cap \mathrm{K}_{ij}(\underline{k})\right)\right), \quad \mathrm{i, j} = 1,...7. \tag{13}$$

is the greatest. The suitable output is $\mathrm{K}_{i\,max, j\,max}(\underline{k})$. $\left(\mathrm{K_p} \cap \mathrm{K_d} \in \{\mathrm{K_{ij}}, i, j = 1,2...7\}\right)$.

So finally the obtained rule-base is:

if e is N & $\dot{e}$ is Z & y is N then kp is Z & kd is P

if e is N & $\dot{e}$ is P & y is Z then kp is Z & kd is Z

if e is Z & $\dot{e}$ is N & y is N then kp is P & kd is Z

if e is Z & $\dot{e}$ is Z & y is Z then kp is Z & kd is Z

if e is Z & $\dot{e}$ is P & y is P then kp is N & kd is Z

if e is P & $\dot{e}$ is N & y is N then kp is Z & kd is Z

if e is P & $\dot{e}$ is Z & y is P then kp is Z & kd is N

if e is P & $\dot{e}$ is P & y is P then kp is Z & kd is P

st4 The inference mechanism is the GMP.

$$\frac{\begin{array}{l}\text{if} \quad \mathrm{E} \cap \dot{\mathrm{E}} \cap \mathrm{Y}(\underline{e}) \quad \text{then} \quad \mathrm{K}(\underline{k}) \\ \quad \mathrm{E_i} \cap \dot{\mathrm{E}}_\mathrm{i} \cap \mathrm{Y_i}(\underline{e})\end{array}}{\mathrm{K_o}(\underline{k})} \tag{14}$$

where $\mathrm{E_i} \cap \dot{\mathrm{E}}_\mathrm{i} \cap \mathrm{Y_i}(\underline{e})$ is the really, actual FLC input.

st5 The defuzzification can be one of the generally accepted methods. The outputs of the j-th rule are $\mathrm{KP}_j^o(Kp)$ and $\mathrm{KD}_j^o(Kd)$, the rule base output is obtained by summarizing all of them:

$$\mathrm{KP}^o(K_p) = \max_{j=1,2...10} \mathrm{KP}_j^o(K_p), \mathrm{KD}^o(K_d) = \max_{j=1,2...10} \mathrm{KD}_j^o(K_d) \tag{15}$$

The FLC outputs after defuzzification are:

$$K_p^* = \frac{\underset{Kp}{sum}\, K_p \cdot \mathrm{KP}^o(K_p)}{\underset{Kp}{sum}\, \mathrm{KP}^o(K_p)}, \; K_d^* = \frac{\underset{Kd}{sum}\, K_d \cdot \mathrm{KD}^o(K_d)}{\underset{Kp}{sum}\, \mathrm{KD}^o(K_d)} \tag{16}$$

The modified system of rules consists of if $\mathrm{E} \cap \dot{\mathrm{E}} \cap \mathrm{Y}(\underline{e})$ then $\mathrm{K}(\underline{k}) \cap \sigma(\underline{k})$

rules.

Thus the output in case of one rule is as follows:

$$\frac{\begin{array}{ll}\text{if} \quad \mathrm{E} \cap \dot{\mathrm{E}} \cap \mathrm{Y}(\underline{e}) & \text{then} \quad \mathrm{K}(\underline{k}) \cap \sigma(\underline{k}) \\ \mathrm{E_i} \cap \dot{\mathrm{E}}_\mathrm{i} \cap \mathrm{Y_i}(\underline{e}) & \end{array}}{\mathrm{K_{poss}}(\underline{k})} \tag{17}$$

$\mathrm{K_{poss}}(\underline{k}) = \mathrm{K_{poss}}\left(K_p, K_d\right)$ is a surface above the K_p,K_d plaine, and it is described with a matrix. The summarized rule base output is computed with *max* too, as in (11).

The defuzzification process is:

$$K_p^* = \frac{\underset{Kp}{sum}\, K_p * \mathrm{KP}^o\left(K_p\right)}{\underset{Kp}{Sum}\, \mathrm{KP}^o\left(K_p\right)}, \quad K_d^* = \frac{\underset{Kd}{sum}\, \mathrm{KD}^o\left(K_d\right) * K_d{}^T}{\underset{Kp}{Sum}\, \mathrm{KD}^o\left(K_d\right)},$$

where the $*$ operation is multiplication of matrixes KP$^\mathrm{o}$, KD$^\mathrm{o}$ and vectors K_p,K_d , and the *Sum* is summa of all the elements of matrixes KP$^\mathrm{o}$, Kd$^\mathrm{o}$.

st6 The defuzzification can be one of the generally accepted methods.

Example 1

Let be $\left[-L_{K_p}, L_{K_p}\right] = \left[-L_{K_d}, L_{K_d}\right] = [-400{,}400]$.

For rule-premise if e is PM and $\dot{e}$ is NM and y is ZE from the rule-base, and for $g(t) = 1 - t$, $q = \infty$ the possibility measure is:

$$\sigma\left(K_p, K_d\right) = g^{(-1)}\left(\frac{\left|K_p \cdot \frac{2}{3} + K_d \cdot \left(-\frac{2}{3}\right) - 0\right|}{\frac{1}{3} \cdot \left(\left|K_p\right| + \left|K_d\right| + 1\right)}\right),$$

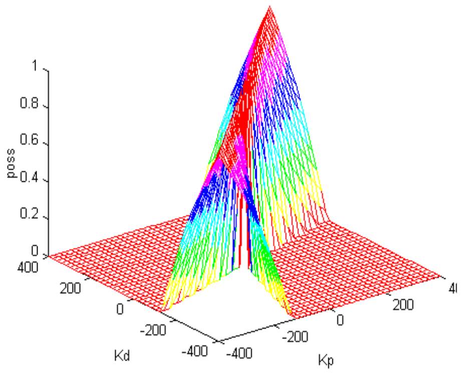


Figure 2

$poss(\, i\,max, j\,max) = poss(4{,}4) = 1$ (see *Figure 2*)

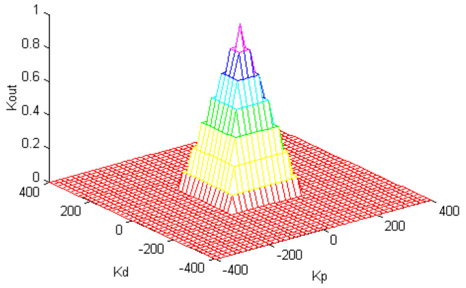


Figure 3

Figure 3 shows the chosen rule-consequence. The complete rule is

if e is PM and $\dot{e}$ is NM and y is ZE then K_p is ZE and K_d is ZE.

3.1 Modified Mamdani Model

The modified model differs from the one described in the previous part, in which instead of using only possibility measure based or only linguistic outputs their intersection is used. Therefore, the modified system of rules consist of if $\mathrm{E} \cap \dot{\mathrm{E}} \cap \mathrm{Y}(\underline{e})$ then $\mathrm{K}(\underline{k}) \cap \sigma(\underline{k})$ rules. Thus the inference mechanism is as follows:

$$\begin{array}{lll} \text{if} & \mathrm{E} \cap \dot{\mathrm{E}} \cap \mathrm{Y}(\underline{e}) & \text{then} \quad \mathrm{K}(\underline{k}) \cap \sigma(\underline{k}) \\ & \mathrm{E_i} \cap \dot{\mathrm{E}}_\mathrm{i} \cap \mathrm{Y_i}(\underline{e}) & \\ \hline & & \mathrm{K_{poss}}(\underline{k}) \end{array}$$

Example 2: For the same parameter-choice from E*xample 1.* $\mathrm{K_{poss}}(\underline{k})$ form is shown on the *Figure 4.* Consequently, the linear dependence of the parameters are not used only in the rule base construction andverification but in the inference mechanism as well, thus narrowing the linguistic rule consequence. Bearing in mind that the rule output is two-dimensional, geometrically the $\mathrm{K_{poss}}(\underline{k})$ forms are more complex nevertheless, a suitable defuzzification procedure can be found.

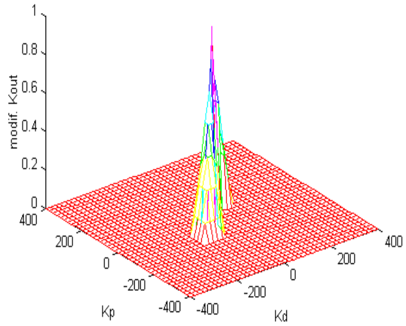


Figure 4

Conclusions

The (1) type law of the PD-type controller, as linear function relationship, was fuzzified using function-fuzzification theory. The calculation of possibility measure offers new horizons for the rule base construction and verification not only in the case of linear function relationship but also in any general function relationships. Out of the values *poss(i max, j max)*, the greatest that determined the $\mathrm{K}(\underline{k})$ domain, is in interval [0,1], and as realization measure of the given rule, it is a rule-weighing. So we obtain a narrowing linguistic rule-consequence. A new method for on-line determination of the gains of a PD controller by using a separate new type fuzzy logic controller is given. Based on the linearity of the control law the possibility measure of the rules of the FLC were introduced. The calculation of these possibility measures offers new horizons for the rule base construction. The proposed new FLC model restricts the Mamdani type rule consequences to possibility domain. In order to verify the performance of the proposed controller simulation has been carried out. It was concluded that in case of the application of generalized t-norms and pseudo-operators in the rules and in the inference mechanism the new method provide better performance than the conventional type controller.

References

[1] R. R. Yager, D. P. Filev, *Essential of FuzzyModeling and Control,*Book, New York/John Wiles and Sons Inc./, 1994

[2] M. Kovács, *An optimum concept for fuzzified mathematical programming problems,*Interactive Fuzzy Optimization, Berlin-Heidelberg-New York/Springer Verlag/, 1991, pp. 36-44

[3] M. Takács, *Fuzzy control of dynamic systems based on possibility and necessity measures,* Proc., RAAD'96, 5th International Workshop on Robotics In Alpe-Adria-Danube Region, Budapest, June 1996

[4] Pap, E., (1997), *Pseudo-analysis as a mathematical base for soft computing*, Soft Computing, 1, pp.61-68

[5] Klement, E. P., Mesiar, R, Pap, E., '*Triangular Norms*', Kluwer Academic Publishers, 2000, ISBN 0-7923-6416-3

[6] Fodor, J., Rubens, M., (1994), *Fuzzy Preference Modeling and Multi-criteria Decision Support*. Kluwer Academic Pub.,1994

Classification of Cerebral Blood Flow Oscillation

Balázs Benyó

Department of Informatics, Széchenyi István University, Egyetem tér 1, H-9026 Győr, Hungary, E-mail: benyo@sze.hu

Péter Somogyi

Department of Control Engineering and Information Technology, Budapest University of Technology and Economics, Magyar tudósok krt. 2, H-1117 Budapest, Hungary, E-mail: psomogyi@bio.iit.bme.hu

Béla Paláncz

Department of Photogrammetry and Geoinformatics, Budapest University of Technology and Economics, Műegyetem rakpart 3, H-1111 Budapest, Hungary, E-mail: palancz@epito.bme.hu

Abstract: Cerebral blood flow (CBF) oscillation is a common feature of several physiological and pathophysiological states of the brains. In this study the characterization of the temporal pattern of the cerebral circulation has been analyzed. The classification of CBF signals has been carried out by two different classification methods – neural network and support vector machine – employing spectral and wavelet analysis as feature extraction techniques. The efficiency of these classification and feature extraction methods are evaluated and compared. Computations were carried out with Mathematica and its Wavelet as well as Neural Networks Applications.

Keywords: Biomedical Systems, Classification, Neural Network models, Radial base function networks, Support Vector Machine

1 Introduction

Low frequency spontaneous oscillations in cerebral hemodynamics have been observed – and linked to certain physiological and patophysiological states [1], such as epilepsy [2]. Therefore it is worthwhile to investigate the possibilities of classification of the temporal patterns of this vasomotion. Three classes of CBF signals have been distinguished experimentally [**3**, 4, 5] in relation to consecutive administration of two different drugs (see Figure 1):

1 Normal blood flow signals before applying any drugs, that does not exhibit low frequency oscillations (LFO-s), referenced as class A;

2 Slight oscillation after the administration of L-NAME, a NO synthase inhibitor reportedly evoking CBF oscillations, referenced as class B;

3 More pronounced oscillation observed after the administration of U-46619 for stimulating thromboxane receptors, having the effect of also inducing LFO, referenced as class C.

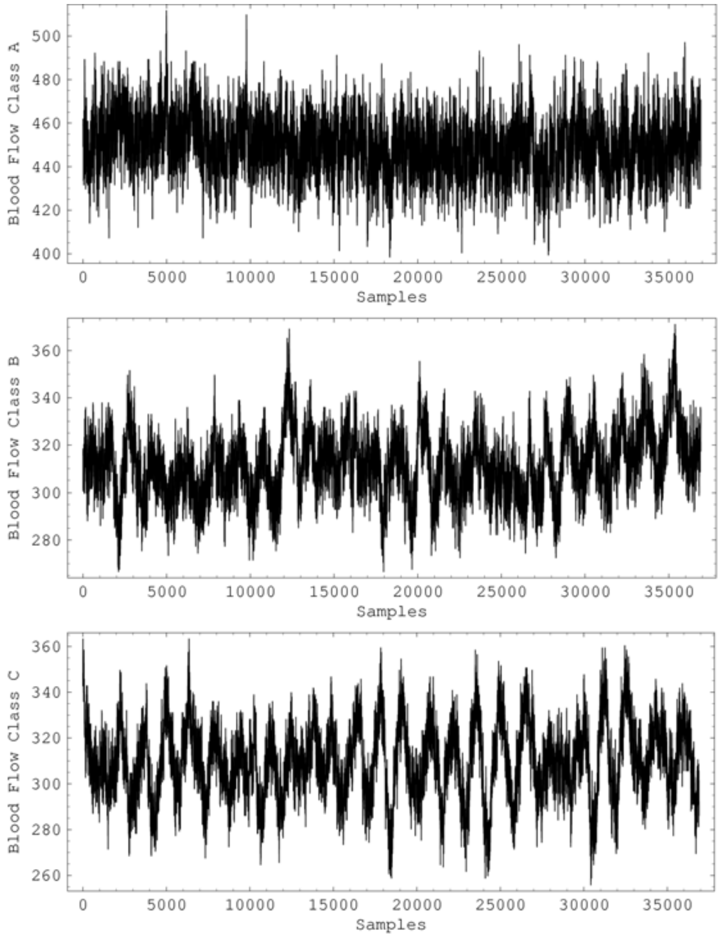


Figure 1
Measured blood flow series of the three states

Recently, to identify different states of CBF oscillation, different classification methods, based on a two-dimensional feature vector – the maximum amplitude and its frequency of the Fourier transform of the time signals – have been employed, using neural network and support vector machine classifiers (SVMC) [6, 7]. However, these approaches were only partly successful because the two-dimensional feature vector could not characterize all the features of the time series. Even the most promising technique, the SVMC suffered from overlearning [8]. The separation of the first class from the two latter has been carried out successfully using two feature vector elements derived from the measured signal. However, the second and third classes cannot be effectively distinguished due to the highly overlapping regions (stars and squares), as seen on the feature map Figure 2. Hence the discrimination of the mentioned classes, or cerebral blood flow states, is the subject of this paper. Two different feature extraction methods have been applied to characterize the given time signals, based on spectral and wavelet subband analysis.

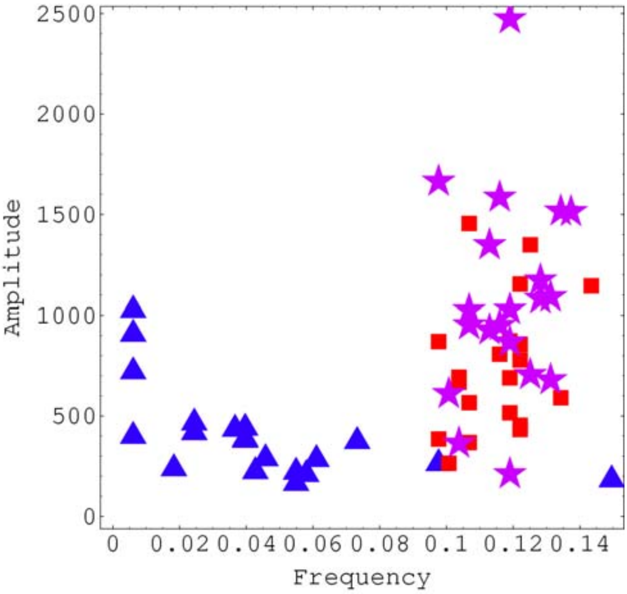


Figure 2

Normalized dimensionless feature map of cerebral blood flow: normal blood flow, class A (triangle), before administration of U-46619, class B (square) and after administration of U-46619, class C (star) from [4]

2 Feature Extraction

2.1 Using Spectral Analysis

In this case, the characterization of the signals to be classified is based on the eigenvalue of the spectral matrix of the signal [9]. In order to obtain the singular values being characteristic of the different states, a matrix has to be derived from the time signal. This is obtained by creating a spectral matrix. Given the time series of data d_i, where $i = [1 \ldots 70.000]$ are the sample points, we pick a window size of $n << 70.000$ and form $70.000 - n$ window vectors, which we apply to a given range of data points:

$$\begin{aligned} &\underline{\mathrm{u}}_1 = (\mathrm{d}_1, \mathrm{d}_2, \mathrm{d}_3, \mathrm{d}_4, \mathrm{d}_5, \mathrm{d}_6, \ldots, \mathrm{d}_\mathrm{n}) \\ &\underline{\mathrm{u}}_2 = (\mathrm{d}_2, \mathrm{d}_3, \mathrm{d}_4, \mathrm{d}_5, \mathrm{d}_6, \mathrm{d}_7, \ldots, \mathrm{d}_\mathrm{n}) \\ &\vdots \\ &\underline{\mathrm{u}}_\mathrm{j} = (\mathrm{d}_\mathrm{j}, \mathrm{d}_{\mathrm{j}+1}, \mathrm{d}_{\mathrm{j}+2}, \ldots, \mathrm{d}_{\mathrm{j}+\mathrm{n}-1}) \end{aligned} \tag{1}$$

The matrix is built from these window vectors as columns:

$$\underline{\underline{\mathrm{A}}} = [\underline{\mathrm{u}}_1^\mathrm{T} \, \underline{\mathrm{u}}_2^\mathrm{T} \cdots \underline{\mathrm{u}}_\mathrm{j}^\mathrm{T}] \tag{2}$$

then our spectral matrix can be computed as $\underline{\underline{S}} = \underline{\underline{A}}^T \underline{\underline{A}}$. In order to find the optimal window size and range, a series of decompositions have been completed, and the reconstructed signals have been compared to the original recordings. A sample of the maximal reconstruction errors can be seen in Figure 3, showing the local minimum and maximum. There is a few percent difference between the local minimum and maximum, therefore for the feature extraction, a window size of 50 and a window range of 5000 samples has been selected.

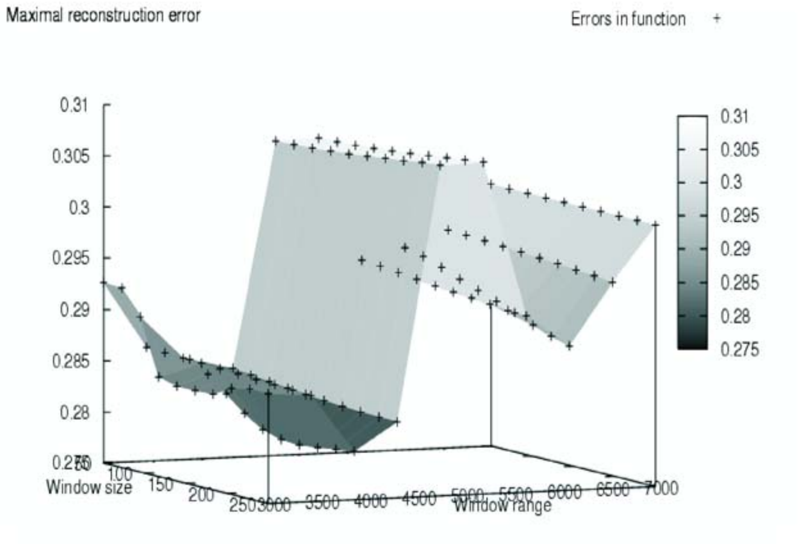


Figure 3
Approximation error in function of window size and range

Employing these window parameters, the eigenvalues of the spectral matrix can be computed. As it can be seen, in the case of a class C signal on Figure 4, the first six values are good candidates to be the elements of the feature vector describing a given signal.

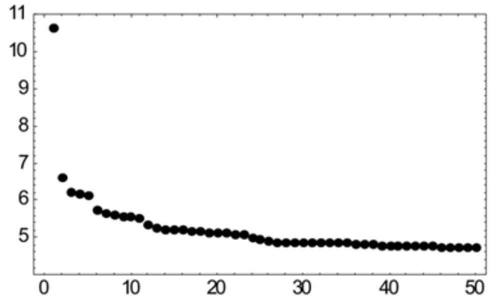


Figure 4
Eigenvalues of the spectral matrix of a class C signal, on a logarithmic scale

2.2 Feature Extraction via Wavelet Transformation

In recent years, feature extraction methods were developed based on wavelet transformation to recognize acoustic signals. They are applicable to the recognition of ships from sonar signatures, cardiopulmonary diagnostics from heart sounds, safety warnings and noise suppression in factories, and recognition of different types of bearing faults in the wheels of railroad cars and so on [10].

Wavelet-based analysis is similar to the Fourier analysis where sinusoids are chosen as the basis function. The Wavelet analysis is also based on a decomposition of a signal using an orthogonal family of basis functions. Because of the used basis function the wavelets are well suited for the analysis of transient, time-varying signals. The wavelet expansion is defined by a two-parameter family of functions:

$$f(t) = \sum_k \sum a_{j,k} \Psi_{j,k}(t) \tag{3}$$

where j and k are integers, the $\Psi_{j,k}(t)$ are the wavelet expansion functions. The wavelet expansion (or basis) functions based on the mother wavelet of the formula

$$\Psi_{j,k}(t) = 2^{j/2}\Psi(2^j t - k) \tag{4}$$

where j is the translation parameter and k is the dilation parameter.

The expansion coefficients $a_{j,k}$ are called the discrete wavelet transform (DWT) coefficients of $f(t)$. The coefficients are given by the following equation

$$a_{j,k} = \int f(t)\Psi_{j,k}(t)dt \tag{5}$$

The DFT and the DWT are the two most commonly used techniques for the signal transformation to the frequency domain. More details of the transforms can be found in [11, 12].

Let us illustrate this classical technique applying it to a CBF signal. Before the DWT of the time signal can be computed, we drop the beginning and the end of this raw signal, getting a signal of 2^{16} length of samples. This transformation decomposes the data into a set of coefficients in the wavelet basis. There are 16 sublists containing the wavelet coefficients in the orthogonal basis of the orthogonal subspaces. The contributions of the coefficients to the signal at different scales are represented by the phase space plot, see Figure 5. Each rectangle is shaded according to the value of the corresponding coefficient: the bigger the absolute value of the coefficient, the darker the area. The time unit is 5 msec.

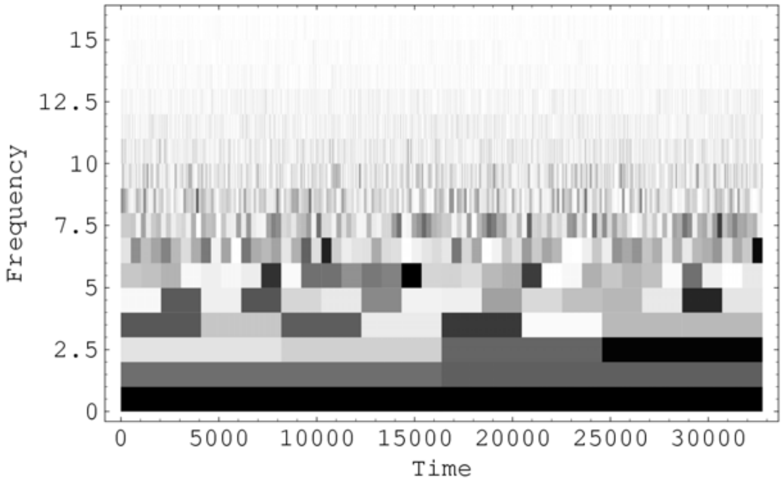


Figure 5
The phase space plot of the DWT of the time signal

Normally, from the wavelet coefficients of each of the 16 resolution levels (subbands) and from sample values of the original time signal, one computes the average energy content of the coefficients at each resolution. There are a total of 17 subbands (16 wavelet subbands and one approximation subband represented by the original signal) from which features are extracted. The i^{th} element of the feature vector is given by,

$$v_i = \frac{1}{n_i} \sum_{j=1}^{n_i} w_{i,j}{}^2, i = 1,2,\ldots,17 \tag{6}$$

where $n_1 = 2, n_2 = 2, n_3 = 2^2, \ldots, n_{16} = 2^{15}$ and $n_{17} = 2^{16}$, where $w_{i,j}$ is the j^{th} coefficient of the i^{th} subband. In this way, from a time signal having 2^k samples or dimensions, one can extract a feature vector of $k+1$ dimensions. This technique has been extended for two dimensional signals, for digital images [13]. In order to study the effect of the dimension of the input space on the quality of the classification as well as to save the morphology of DWT, here we employ a different approach. We consider the wavelet coefficients belonging to a given subband as a feature vector based on this given resolution. It can be a reasonable approach, because the approximated signal representation in the orthogonal subspace corresponding to this subband is given by these coefficients [14]. In our case, there are two sets of time signals, representing two classes of CBF states and only 40 patterns (2×20) are at our disposal. Intuitively, it is possible to shatter two points by any linear manner in the one-dimensional space and three points in two-dimensional space. By analogy, it is possible to shatter $\mathrm{N}+1$ points in the N-dimensional space with the probability of 1. If the patterns to be classified are independent and identical distributed, then in the 2 N patterns are linearly separable in the N-dimensional space [15]. The coefficients of the subbands from $n_2 = 2$ up to $n_6 = 2^5 = 32$ as different feature vector components will be employed. The magnitudes of the wavelet coefficients at these subbands are shown in Figure 6.

To carry out the computations, a second order Daubechies filter has been employed, see Figure 7.

According to [6], the two greatest components of a wavelet decomposition do not represent adequately the signals derived from drug induced oscillations. This means that the very small coefficients belonging to the higher resolutions, $n_1 - n_{16}$ and the very big coefficients of the lowest resolution, (n_1) are not taken into consideration. The previous ones have no contributions; the latest one would suppress all of the others (see Figure 5). With other words, we consider the "measurable" fine structure of the subband coefficients. Figure 8 shows the maximums of the magnitude of the wavelet coefficients of different resolutions, except of those belonging to the first (lowest) one. The omitted first wavelet coefficient has a magnitude of about 74276, being significantly larger than the other wavelet coefficients.

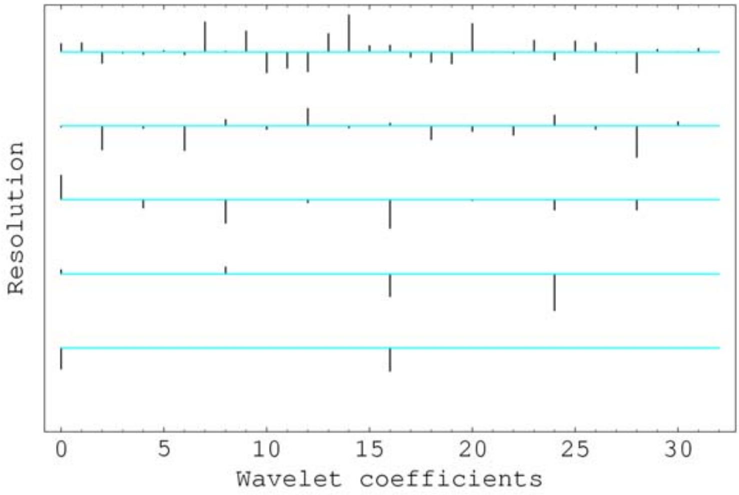


Figure 6
The magnitude of the wavelet coefficients at resolution from (at the bottom) up to (at the top)

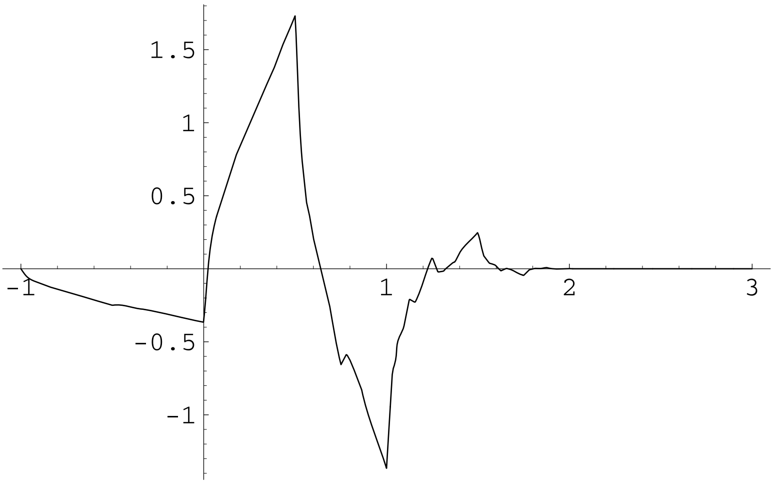


Figure 7
Daubechies filter with $\psi(t)$

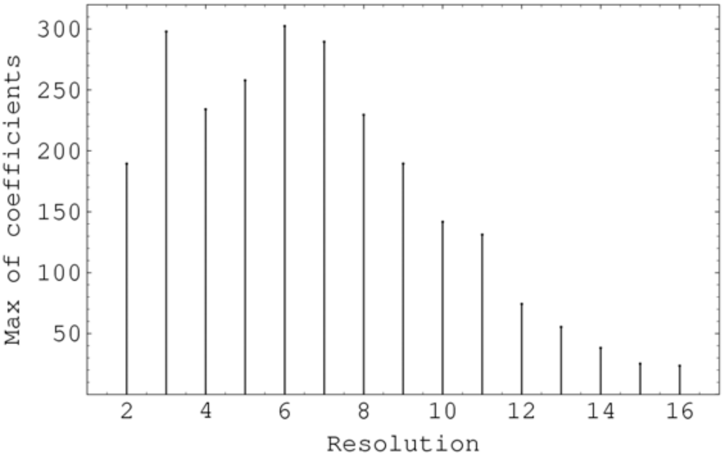


Figure 8
The maximal magnitudes of the wavelet coefficients of different resolutions

3 Classification

3.1 Using Radial Basis Function with Artificial Neural Networks

Figure 9 illustrates an RBF network with inputs $x_1,\ldots,x_n$ and output $\hat{y}$. The arrows in the figure symbolize parameters in the network. The RBF network consists of one hiddenlayer of basis functions, or neurons. At the input of each neuron, the distance between the neuron centre and the input vector is calculated. The output of the neuron is then formed by applying the basisfunction to this distance. The RBF network output is formed by a weighted sum of the neuron outputs and the unity bias shown.

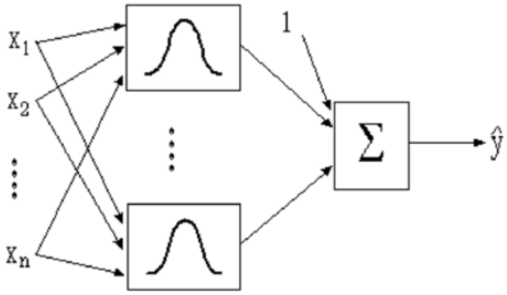


Figure 9
Illustration of an RBF network

The RBF network in Figure 9 is often complemented with a linear part. This corresponds to additional direct connections from the inputs to the output neuron.

Mathematically, the RBF network, including a linear part, produces an output given by

$$\hat{y}(\theta) = g(\theta,x) = \sum_{i=1}^{nb} w_i^2 e^{-\lambda_i^2(x-w_i^1)^2} + w_{nb+1}^2 + \chi_1 x_1 + \ldots + \chi_n x_n \tag{7}$$

where nb is the number of neurons, each containing a basis function. The parameters of the RBF network consist of the positions of the basis function w_i^1, the inverse of the width of the basis functions λ_i, the weights in output sum w_i^2 and the parameters of the linear part $\chi_1 x_1 + \ldots + \chi_n x_n$. In most cases of function approximation, it is advantageous to have the additional linear part but it can be excluded when not necessary.

Considering N patterns of measured CBF signals representing the two overlapping classes, we have

$$x_i \in R^M \tag{8}$$

feature vectors derived from time series samples, where $i = 1\ldots N$ are the samples, and M is the dimension of the feature vectors, consisting of first M dominant eigenvalues. In our case the number of the measurements were $N = 40$. In order to obtain the minimum size of the feature vector which is required to produce reliable results, up to six eigenvalues were used. The goal of the classification problem is to assign new, previously unseen patterns to their respective classes based on previously known examples: in our case to assign input signals to class B or class C. Therefore the output of our unsupervised learning algorithm is a set of discrete class labels corresponding to the different CBF states. The labelled patterns corresponding to classes B and C, were to be classified. This means, that we are looking for a decision function; the output of this estimating function is interpreted as being proportional to the probability that the input belongs to the corresponding class. To carry out the systematic classification of CBF signals, an RBF was used. Radial basis function has the characteristic feature, that the response increases or decreases monotonically according to the distance from a central point. The Gaussian RBF function is used in a single layer network, consisting of two input nodes in the input layer, seven nodes in the hidden layer, and two nodes in the output layer. Several lengths of feature vectors have been fed to the classifier – producing fewer misclassifications as the number of components of the input feature vector increased. The output of the classifier was accepted, if the rounded value of the output nodes corresponded to the proper class, correctly classifying the input pattern, otherwise the classification of that particular input signal was registered as a misclassification. Because the number of the samples are limited ($N = 40$), the dimension of the feature vectors, as well as the number of nodes of the network should be constrained, in order to ensure a reliable teaching process of the network.

3.2 Support Vector Machine (SVM) Classifier

To study the effects of higher dimensional feature vectors on the classification process, an SVM classifier, having no restrictions regarding input vector dimensions, has also been applied.

Let us consider a set of training samples,

$$S = ((x_1, y_1), \ldots, (x_m, y_m)) \tag{9}$$

with labels $y_i = 1$ or -1, respectively and use the feature space implicitly defined by the kernel $K(x,z)$. We suppose that the parameters are the solutions of the following quadratic optimization problem,

$$\text{maximize } W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j \left(K(x_i, x_j) + \frac{1}{c} \delta_{i,j} \right) \tag{10}$$

$$\text{subject to } \sum_{i=1}^{m} y_i \alpha_i = 0, \alpha_i \geq 0, i = 1 \ldots m \tag{11}$$

$$\text{Let } f(x) = \sum_{i=1}^{m} y_i \alpha_i^* K(x_i, x) + b^* \tag{12}$$

where b^* is chosen so that

$$y_i f(x_i) = 1 - \frac{\alpha_i^*}{c} \tag{13}$$

for any i with

$$\alpha_i^* \neq 0 \tag{14}$$

Then the decision rule given by $sign(f(x))$ is equivalent to the hyperplane in the feature space implicitly defined by the kernel $K(x,z)$, which solves the optimization problem, where the geometric margin is

$$\gamma = \left(\sum_{i \in sv} \alpha_i^* - \frac{1}{c} \left\langle \alpha^*, \alpha^* \right\rangle \right)^{-\frac{1}{2}} \tag{15}$$

and the set sv corresponds to indexes i, for which $\alpha_i^* \neq 0$,

$$sv = \{ i : \alpha_i^* \neq 0; i = 1 \ldots m \} \tag{16}$$

Training samples, x_i, for which $i \in sv$ are called support vectors contributing to the definition of $f(x)$. The geometric margin, γ can indicate the quality of the classification [16], greater the γ, more reliable the classification is.

This kernel based classifier can be trained on any size of training set, while neural networks should have so many input nodes as the dimension of the input space and need definitely more training patterns than the number of these input nodes. Employing kernels, a classification problem can be transferred in a higher dimensional space, where the linear separability is more likely. In addition, the quality of the classification in any dimension can be measured by the geometric margin of the SVM classifier [16]. Here we used the feature vectors produced by the wavelet subband analysis. Twenty of these vectors represent one CBF state, the other twenty represent the other state. As an example let us load the coefficients of the fifth subband, $n_5 = 2^4 = 16$, for all of the 40 patterns, giving us 40 feature vectors of dimension of 16. First, these data should be standardized; to be transformed so that their mean is zero and their unbiased estimate of variance is unity. Let us employ Gaussian kernel, with parameter $\beta = 5$, as seen on Figure 10.

$$K(u,v) = e^{-\beta(u-v)^T(u,v)} \tag{17}$$

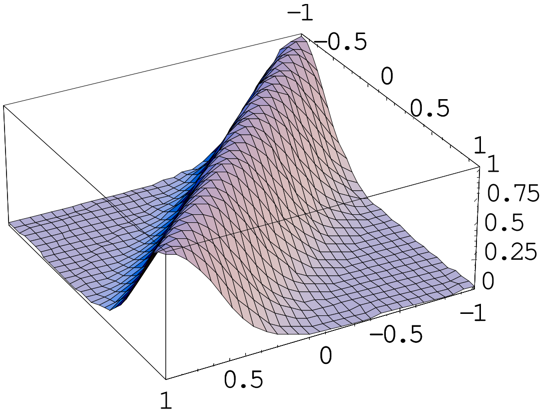


Figure 10

Gaussian RBF universal kernel with $\beta = 5$, in case of $x, y \in R^1$

Let the value for the control parameter of regularization be $c = 100$. To carry out the training of the support vector classifier, we shall employ the algorithm embedded into the function, *SupportVectorClassifier* developed for *Mathematica* [17]. A sample pattern can be considered as support vector, if its contribution (its weighting coefficient α_i) to the decision function is greater than 1% of the maximal contribution.

These computations were carried out for different feature vectors based on the coefficients of the different subbands.

4 Results

4.1 Result of the SVM Classification

Table 1 shows, that by decreasing the number of the wavelet coefficients, the Gram matrix is getting ill-conditioned, the geometric margin is becoming narrower and probability of the misclassification of patterns is increasing, although the classification with four wavelet coefficients is just acceptable. Let us employ the traditional feature extraction method, when the elements of the feature vector are computed as the average of squares of the wavelet coefficients belonging to the same subband, plus the same contribution of the original signal as additional subband. Consequently, the dimension of the feature vector is $16+1=17$. Table 2 shows the result for this case. These results correspond with the results of the classification carried out with the eight dimensional feature vectors based on subband level 4, however now the dimension of the feature vectors is 17 instead of 8. The robustness of the SVM classifier has been also proved by successful classification of noisy samples.

Table 1

The results of the SVM classification with different feature vectors

Subband level	Number of wavelet coefficirnts	Determinant of Gram matrix	Condition number of Gram matrix	Number of support vectors	Geometric margin	Number of misclassified patterns
6	32	1.	1.	40	0.159695	0
5	16	1.	1.	40	0.159695	0
4	8	0.999	1.040	40	0.159701	0
3	4	0.005	69.374	40	0.113922	0
2	2	1.9310^{-39}	1.1510^{7}	25	0.083355	4

Table 2

The results of the SVM classification employing traditional feature extraction technique

Determinant of Gram matrix	Condition number of Gram matrix	Number of support vectors	Geometric margin	Number of misclassified patterns
0.994	1.170	40	0.159374	0

4.2 Result of the ANN Classification

Columns 1 and 2 of Table 3 show the result of the ANN classification results using eigenvalue-based feature extraction. It can be clearly seen from the numbers, that it makes no sense to use more than 6 eigenvalues. Comparing different feature extraction methods and classification algorithms by taking different numbers of eigenvalues as feature vectors, the results are very close to that obtained when using wavelet decomposition, see columns 3 and 4 of Table 3. In any case, it is clear, that a merely two element feature vector is insufficient for reliable results; at least a five element feature vector was needed to differentiate class B from class C in the case of ANN classification with eigenvalue-based feature extraction, while in the case of the SVM classification with wavelet-based feature extraction, at least 8-dimensional feature vector should be used.

Table 3

Misclassification rate

Wavelet coefficient (subband level)	Number of eigenvalues	Wavelet feature extraction & SVM classification, misclassification number	Eigenvalue feature extraction & ANN classification, misclassification number
16 (5)	6	0	0
8 (4)	5	0	0
4 (3)	4	0	3
-	3	-	3
2 (2)	2	4	6

Conclusions

Two feature extraction and classification methods are presented. First an Artificial Neural Network using a radial based function, combined with a spectral matrix based feature extraction was shown. Secondly, a Support Vector Machine Classifier with wavelet subband analysis as feature extraction method was employed. The two methods can successfully differentiate cerebral blood flow classes B and C, and although the approaches described in this paper are very different, they still produced comparable results for this classification problem.

Acknowledgement

The research was supported by the Hungarian National Research Fund, Grants No. OTKA F046726, T042990, and by the Hungarian Ministry of Economics and Transportation Grant No. RET-04-2004.

References

[1] H. Nilsson and C. Aalkjær, “Vasomotion mechanisms and physiological importance.”, Molecular Interventions, 7:59-68, 2003

[2] e. a. B. Diehl, “Spontaneous oscillations in cerebral blood flow velocities in middle cerebral arteries in control subjects and patients with epilepsy.”, Stroke, 28:2457-2459, 1997

[3] e. a. G. Lenzsér, “Nitric oxide synthase blockade sensitizes the cerebrocortical circulation to thromboxane-induced CBF oscillations.”, Journal of Cerebral Blood Flow and Metabolism, 23:88, 2003

[4] e. a. Z. Lacza, “The cerebrocortical microcirculatory effect of nitric oxide synthase blockade is dependent upon baseline red blood cell flow in the rat.”, Neuroscience Letters, 291:65-68, 2000

[5] e. a. Z. Lacza, “NO synthase blockade induces chaotic cerebral vasomotion via activation of thromboxane receptors.”, Stroke, 32:2609-2614, 2001

[6] e. a. B. Benyó, “Characterization of the Temporal Pattern of Cerebral Blood Flow Oscillations”, In Proc. of 2004 International Joint Conference on Neural Networks, pp. 468-471, July 2004

[7] e. a. B. Benyó, “Classification of Cerebral Flood Flow Oscillation using SVM classifiers with different kernels”, Intelligent Systems at the Service of Mankind, 2:145-151, 2005

[8] e. a. B. Benyó, “Characterization of Cerebral Blood Flow Oscillations Using Different Classification Methods”, In Proc. of 2005 IFAC, 2005. electronic publication #04987

[9] W. Shaw and J. Tigg, ”Applied Mathematica”, Addison-Wesley, 1994

[10] J. Goswami and A. K. Chan, ”Fundamentals of Wavelets. Theory, Algorithms, and Application”, Wiley, 1999

[11] e. a. Li, “Detection of ECG Characteristic Points Using Wavelet Transforms”, IEEE Trans. on Biomedical Eng., 1995, Vol. 42, pp. 21-28

[12] e. a. L. Senhadji, “Comparing Wavelet Transforms for Recognizing Cardiac Patterns”, IEEE EMBS Magazine, Vol. 14, pp. 167-173

[13] B. Paláncz, “Wavelets and their Application to Digital Image Processing”, Publications in Geomatics, 2004, Győr, ISSN 1419-6492

[14] Aboufadel and Schlicker, ”Discovering Wavelets”, Wiley-Interscience, 1999

[15] e. a. L. Zhang, “Wavelet Support Vector Machine”, IEEE Trans. Systems, Man and Cybernetics - Part B: Cybernetics, 4(1):34–39, Februray 2004

[16] N. Cristianini and J. Shawe-Taylor, ”An introduction to Support Vector Machines and other kernel-based learning methods”, Cambridge University Press, 2000

[17] B. Paláncz, “Support Vector Classifier via Mathematica”, Wolfram Research Mathematica Information Center, 2004, http://library.wolfram.com/infocenter/mathsource/5293