A genetic based approach to the Type I structure identification problem

S.E. Papadakis,

Technological Educational Institution of Kavala, Dept. of Industrial Informatics, Div. of Computing Systems, GR-65404 Kavala, Greece. E-mail: steliospapadakis@mail.gr

P. Tzionas,

Technological Educational Institution of Thessaloniki, Dept. of Automation, Thessaloniki, Greece. E-mail: ptzionas@teithe.gr

V.G. Kaburlasos,

Technological Educational Institution of Kavala, Dept. of Industrial Informatics, Div. of Computing Systems, GR-65404 Kavala, Greece. E-mail: vgkabs@teikav.edu.gr

J.B. Theocharis,

Aristotle University of Thessaloniki, Dept. of Electrical & Computer Engineering, Thessaloniki, Greece. E-mail: theochar@vergina.eng.auth.gr

Abstract. The problem of system input selection, dubbed in the literature as Type I Structure Identification problem, is addressed in this paper using an effective novel method. More specifically, the fuzzy curve technique, introduced by Lin and Cunningham (Lin 1995), is extended to an advantageous fuzzy surface technique; the latter is used for fast building a coarse model of the system from a subset of the initial candidate inputs. A simple genetic algorithm, enhanced with a local search operator, is used for finding an optimal subset of necessary and sufficient inputs by considering jointly more than one inputs. Extensive simulation results on both artificial data and real world data have demonstrated comparatively the advantages of the proposed method.

Key words: structure identification, feature selection, modeling, input selection, fuzzy curve, fuzzy surface, genetic algorithms.

1. Introduction. System identification from input-output data pairs has always been a challenging topic in engineering. Recently, computational intelligence techniques, including neural networks and neuro-fuzzy models have been successfully used, as universal model free estimators (Lee 1990, Lin 1992, Wang 1992a, Wang 1992b) for modeling, identification and control of ill-defined real world processes. Despite the large number of published works on novel cognitive models for modeling complex physical systems in a black-box fashion, the computation of a *convenient structure* for a model, i.e. a model with both significant identification performance and low complexity, remains active. Sugeno and Yasukawa (Sugeno 1993) have dealt with the structure identification problem by

dividing it into two other (sub)problems namely Type I and Type II problems. The latter problem, which refers to both the architecture of a model and the parameter identification (training), is beyond the scope of this paper. The Type I problem refers to the selection of those inputs (or, equivalently, features in classification literature) that affect the output of the system substantially. According to the authors in (Sugeno 1993) the Type I structure identification problem is one hundred times more important than the Type II structure identification problem. Moreover, the Type I problem is subdivided into Type Ia and Type Ib (sub)problems. More specifically, the Type Ia problem concerns the ad-hoc definition of a set of candidate inputs. This work focuses on the Type Ib structure identification problem, that is the selection of an optimal subset of inputs, which are necessary and sufficient for describing the system.

The objective of input selection is to determine a feature space with a) low dimensionality, b) retention of sufficient information, c) enhancement of separability of the feature space, for example in different categories by removing effects due to noisy features, and d) comparability of features among examples in same category (Piramuthu 2004).

Various methods have been proposed in the literature for system input selection including statistical (Fukunaga 1990, Kittler 1975), geometrical information-theoretic measures (Battiti 1994), mathematical programming (Bradley 1998) methods, etc. More specifically, in statistical analysis, forward and backward stepwise multiple regression (SMR) are widely used to select features, with forward SMR used more often due to fewer calculations. In the latter case the output is the smallest subset of features resulting in a correlation coefficient value that explains a significantly large amount of the variance. Similarly, in (Malki 1991), the K-L transform was applied to the training examples in order to obtain the initial training vectors. Training is started in the direction of the major eigenvectors of the correlation matrix of the training examples. The main drawback of the aforementioned methods is that they seek input interdependencies exclusively in the input space, ignoring the influence of each input to the output of the system.

In (Siedlecki 1998) genetic algorithms were used for feature selection by encoding an initial set of n features as n-element bit string with 1 and 0 representing the presence and absence, respectively, of features in the set. The authors used classification accuracy as the fitness function (for genetic algorithms, while selecting features) and obtained good neural network results compared to branch-and-bound and sequential search algorithms. However, latter work (Hopkins 1994) was shown that classification accuracy may be a poor fitness function measure when seeking to reduce the dimension of the feature set. Also, the time complexity of the method is overwhelming due to the training process required for each input combination.

Rough sets theory was also used to determine the degree of dependency of sets of attributes for selecting binary features. Features leading to a minimal preset decision tree, which is the one with the minimal length of all paths from root to

leaves, were selected. Best first search was also used stopping after a user defined number of non-improving node expansions (Piramuthu 2004). Similar algorithms including the IDG algorithm use the positions of examples in the instance space to select features for decision trees; they limit their attention to boundaries separating examples belonging to different classes, while rewarding (penalizing) rules that separate examples from different (same) classes (Piramuthu 2004). Decision trees generated using the proposed algorithm had better accuracy. Nevertheless, an inherent drawback of this hierarchical approach is that the structure of a decision tree depends on the specific sequence of features to be tested, apart from the features themselves. Neural networks were also used to measure the contribution of individual inputs to the output of the neural network (Piramuthu 2004). These methodologies have to undergo the time-consuming training process of the network used to test every input combination.

The most popular input selection methods in machine learning literature are variations of sequential forward search (SFS) and sequential backward search (SBS) (Piramuthu 2004). SFS (SBS) obtains a chain of nested subsets of features by adding (subtracting) the locally best (worst) feature in the set. The serious weakness of this approach is that it adds or subtracts one feature at a time, hence it may result in getting trapped in local minima because it fails to encode all possible combinations.

In (Lin 1995) Lin and Cuningham proposed a very fast method for input selection introducing the fuzzy curve technique. More specifically, a fuzzy curve is a non-linear continuous curve, which establishes a connection between a specific input and the output, performing a projection of the multidimensional input output space on the (probed input)- output space. The height of the projected output is the measure of importance of the corresponding input as follows: On the one hand, if the height is sizeable the respective input is considered significant. On the other hand, a substantial input results in a flat fuzzy curve. The basic advantage of this approach is the linear time complexity of the probing process in the number of inputs. The serious weakness of the method arises from the fact that during probing a particular input the rest ones are ignored. Hence, a specific input could be rejected if probed alone, although the same input could be significant if considered jointly with another one.

In this work we use a simple genetic algorithm for feature selection by exploring simultaneously different combinations of inputs. To cope with the above mentioned weakness, we introduce the fuzzy surface concept as an extension of the fuzzy curve. The fuzzy surface is used for developing a coarse model for each input combination. Evaluation of the models is performed by a fitness function combining both the learning performance and the size of the selected input subset. In conclusion, the minimum input subset is identified as that which provides low dimensionality feature space while retaining sufficient information as described here.

The rest of the paper is organized as follows. Section 2 details the proposed fuzzy surface method. In section 3 we present the genetic algorithm used to select the optimal input combination. Section 4 presents experimental application results in three artificial examples and three real word problems. Finally, section 5 summarizes the contribution of this work including plans for future work.

2. The proposed method. Consider a m-input single output system described by a nonlinear function of the form: $y = f(\underline{x})$, where $\underline{x} = [x_1, ..., x_j, ..., x_m]^T$ is the input vector and y denotes the output of the system. Also, let $\wp_{q,m,1}$ denote the observed input/output data set comprising q m-input/output observed patterns: $\wp_{q,m,1} = \left\{ (\underline{x}^k, y^k), \ k = 1, ..., q \right\}$. Let $\Im_{m,1} = \{x_1, x_2, ..., x_j, ..., x_m\}$ be the set of candidate inputs. Also let $\Im_{n,\ell} \mid 1 \leq n \leq m, \ \ell = 1, ..., m!/(m-n)!$ be a particular subset of $\Im_{m,1}$ comprising n of m inputs. The number n represents the cardinality of $\Im_{n,\ell}$. The index ℓ denotes a specific input subset of order n. There are $\dfrac{m!}{(m-n)!}$ subsets of order n. The total number of subsets of any order derived from the set of m candidate inputs, excluding the empty subset, equals $2^m - 1$. Given a specific subset $\Im_{n,\ell}$ the corresponding $\wp_{q,m,\ell}$ of the specific n-inputs-output data observed data points is a subset of $\wp_{q,m,1}$. For each n-input-output datum in $\wp_{q,n,\ell}$ a fuzzy rule with crisp output is created in the following form:

$$R^k : if \ x_1 \ is \ A_1^k \ and \ x_2 \ is \ A_2^k \ and...and \ x_n \ is \ A_n^k \ then \ y \ is \ y^k \tag{1}$$

where the membership function $\mu_j^k(x_j)$ of a fuzzy set $A_j^k \mid j = 1, ..., n$ and $k \in \{1, ..., q\}$ is given by:

$$\mu_j^k(x_j) = \exp[-(\frac{x_j^k - x_j}{\sigma_j})^2] \tag{2}$$

Each bell-shaped function is located at x_j^k; the parameter σ_j has a fixed value for each input x_j, equal to 5-15% of the x_j variable range. Hence, a fuzzy rule base is generated comprising q rules, R^k, $k = 1, ..., q$ in the form of equation (1). Having determined the product as the fuzzy implication method and using the centroid defuzzification technique, the output of the fuzzy model is given by the formula:

$$y = FS_{n,\ell}(\underline{x}) = \frac{\sum\limits_{k=1}^{q} \left[\prod\limits_{j=1}^{n} (\mu_j^k(x_j^k)) \right] \cdot y^k}{\sum\limits_{k=1}^{q} \left[\prod\limits_{j=1}^{n} \mu_j^k(x_j) \right]} \tag{3}$$

Equation (3), which provides a continuous and parameter free surface, is used as a fuzzy model approximating the input-output data.

The mean absolute percentage error is used to estimate the quality of the approximation:

$$E_{q,n,\ell} = \frac{100}{q} \sum_{k=1}^{q} \frac{\left| FS_{n,\ell}(\underline{x}^k) - y^k \right|}{|y^k|} \% \tag{4}$$

To overcome the risk of overfitting due to the large number of rules, the data set is subdivided into two subsets, each consists of the one half ($q/2$) of input/output patterns. The fuzzy surface is built using the first data subset $\wp_{q/2,n,\ell}$ and it is valuated on the whole data set $\wp_{q,n,\ell}$ according to (4).

During the last years, the fuzzy modeling approach has gained considerable interest for solving real world problems, including modeling of highly complex systems, signal processing and pattern recognition (Siedlecki 1998, Hopkins 1994, Lin 1995, Wang 1995, Papadakis 2002, Piramuthu 2004). Extensive experimentation has demonstrated that fuzzy systems exhibit a number of significant advantages compared to other artificial intelligence models, such as the neural network models (Hopkins 1994, Wang 1995, Papadakis 2002). First of all, the neural networks are global models where training is performed on the entire pattern range. On the contrary, owing to the partition of the input space, the fuzzy models perform a fuzzy blending of local models. As a result, faster convergence is achieved. Secondly, fuzzy neural networks are capable of incorporating both numerical data (quantitative information) and expert knowledge (qualitative information) and describe them in the form of linguistic IF-THEN rules. In that respect, they provide a unified framework for integrating the computational parallelism and low-level learning of neural networks with the high-level reasoning of fuzzy systems. The above feature may assist in determining the initial structure, also leading to models with fewer parameters when compared to neural networks.

As regards the representation capabilities, it has been verified based on analytical aspects that fuzzy models are free function approximators (Wang 1992a). To this end, fuzzy models can be employed to identify a broad class of systems described in terms of a general non-linear functional expansion including lagged inputs and outputs. Let us consider a discrete-time dynamical system of this type, represented by:

$$y(t) = f\left[y(t-1), y(t-2), \ldots, y(t-n_y), u(t), u(t-1), \ldots, u(t-n_u)\right] + e(t) \tag{5}$$

where $y(t)$ is the output of the system at time instant t, $u(t-i), i = 1, \ldots, n_u$ are the external inputs at previous time steps, and n_y, n_u are the maximum lags for the output and the input terms, respectively. Furthermore, $e(t)$ is the modeling error regarded as additive white noise, and $f(\cdot)$ is a functional form representing the input-output mapping. The above description, which belongs to a general system formulation, is referred to the literature as the non-linear au-

toregressive moving average with exogenous inputs (NARMAX) (Chen 1989). This is an extension to the nonlinear framework of the traditional linear model regression analysis (Box 1970). It is rigorously proved (Leontaritis 1985) that a discrete-time, time-invariant system can be described by means of the model formalism in (5), provided that two sufficient conditions are valid, namely, the response function $f[\cdot]$ is finitely realizable, and a realizable model exists in the vicinity of the chosen equilibrium point.

It should be noticed that for a practical system, the functional form $f[\cdot]$ is very complicated and usually, not available in advance. Hence, a model needs to be devised to implement the input-output mapping. In this paper, the fuzzy model (3) is used to model the non-linear system described by (5). Initiating the modeling procedure, we define a state vector that is composed of past values of the external input and the output of the system as:

$$\underline{x} = [y(t-1), \ldots, y(t-n_y),\, u(t), \ldots, u(t-n_u)]^T \qquad (6)$$

The state vector is formulated through the use of tapped delay lines, with the state components regarded as inputs to the fuzzy model. Following a series-parallel approach (Feng 2002) shown in Fig. 1, the system is approximated by

$$y(t) = \hat{f}[\underline{x}] + \varepsilon(t) \qquad (7)$$

where $\hat{f}[\underline{x}]$ is an approximation of the actual system functional, a task accomplished by means of the fuzzy model $FS_{n,\ell}(\underline{x})$, and $\varepsilon(t)$ is the residual error. The identification process involves two major issues that should be properly addressed. The first issue regards the selection of the model inputs and it is the primary focus of the present paper. Since the lag orders are generally unknown, a selection mechanism has to be developed, so that among the pool of past input and output values, the most significant ones are chosen as model inputs. The second issue to be tackled relates to the construction of the fuzzy model. This consists of determining the fuzzy partition of the input space, that is, define the number of the fuzzy rules and the location of the membership functions. Additionally, the parameter values at the consequent part of the rules should be computed, based on the available input-output data.

It should be noticed that in our input selection approach, a separate fuzzy model is generated for each subset of inputs of the candidate inputs set, considered in the genetic algorithm, to be presented in the sequel. In that respect, since the modeling task is embedded within the input combination search, we follow the simplified modeling technique discussed above. Owing to the increased complexity and the large number of fuzzy rules included in the formulation, the aforementioned model cannot be used as a regular model in order to identify the system. Nevertheless, it serves as means to built a coarse model for each input combination. Thus we avoid any time-consuming adaptation procedure, reducing the overall computational burden considerably.
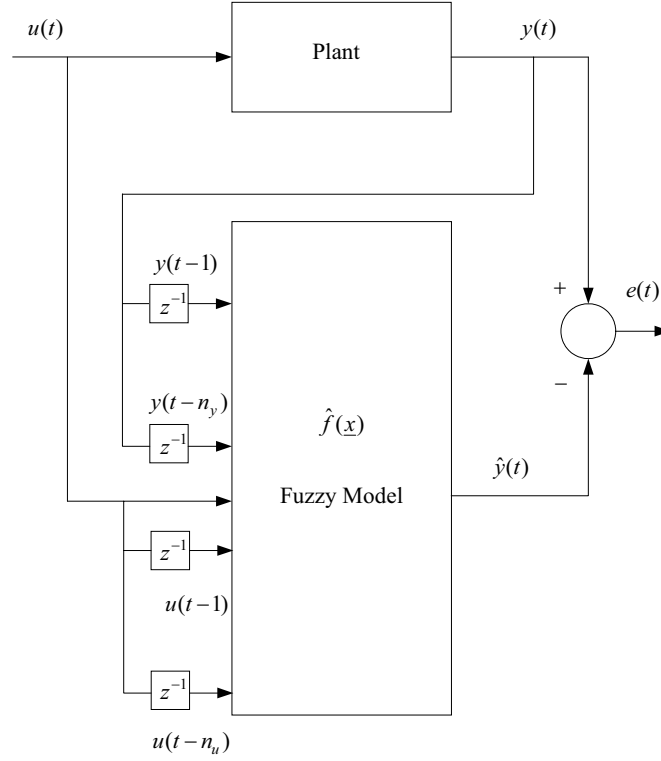
Fig. 1.   Series-parallel configuration for the identification of dynamical systems using the suggested fuzzy model.

Parameters σ_j play a very important role in the proper identification of the system as follows: On the one hand, if the values of the parameters are extremely small the fuzzy surface-based model (FS) results in over fitting the data, increasing $E_{q,n,\ell}$ due to poor generalization. On the other hand, very large value of the parameters σ_j result in inadequate identification performance, increasing the $E_{q,n,\ell}$ due to poor identification. Empirical studies have shown that a value calculated by the formula:

$$\sigma_j = \delta \cdot \left(\max(x_j^k) - \min(x_j^k) \right), k = 1, ..., q \mid \delta \in [0.05, 0.15] \qquad (8)$$

is an appropriate option. Moreover, this calculation is adopted in (Lin 1995). Henceforth, all values of the tunable parameters are calculated by formula (8) overcoming any iterative time-consuming processes.

3. The genetic based input selection algorithm. For each subset $\Im_{n,\ell}$ of inputs a fuzzy surface $FS_{n,\ell}$ can be built using the first $\wp_{q/2,n,\ell}$ input data, according to equation (3). For each fuzzy surface $FS_{n,\ell}$ an evaluation measure

$R_{n,\ell}$, related both to the modeling performance and to the complexity of $FS_{n,\ell}$, is computed by the formula:

$$R_{n,\ell} = w \cdot \frac{100 \cdot E_{q,n,\ell}}{E_{\max}} + (1-w) \cdot \frac{n}{m} \mid w \in (0,1) \qquad (9)$$

where the $E_{\max} = \max\{E_{q,1,\ell}\}_{\ell=1,...,m}$, is used as normalization factor of $E_{q,n,\ell}$ values. The smaller the value of $R_{n,\ell}$ the greater is the importance of $FS_{n,\ell}$ and the associated input combination. If a particular subset $\Im_{n,\ell}$ includes non-important inputs, then both terms in (9) increase, raising the respective $R_{n,\ell}$ value. Inserting significant inputs or discarding insignificant ones the respective terms in (9) decrease resulting in smaller $R_{n,\ell}$ values. In the case that two input combinations are equivalent in the first term of (9), the preferred one is that with the lowest order n.

The objective of the proposed input selection method is to track down the subset $\Im_{n,\ell}$ with the minimum $R_{n,\ell}$ value among all $2^m - 1$ subsets of candidate's inputs $\Im_m$. In case the cardinality m of $\Im_m$ is small probing all possible combinations is feasible. Nevertheless, for a large set of candidate inputs the number of combinations to be probed is prohibitively high. The minimization of $R_{n,\ell}$ can be formulated as a combinatorial unconstrained optimization problem, treated here by an enhanced genetic algorithm with binary encoding, adaptive crossover/mutation rates and elitism (Papadakis 1996). More specifically, the chromosome of each individual consists of m genes of one bit. Each bit $b_j \in \{b_1, b_2, ..., b_m\}$ encodes either the presence '1' or the absence '0' of its refereed input variable $x_j \mid j = 1, ..., m$ thus creating a particular $\Im_{n,\ell} \subseteq \Im_{m,1}$. Hence, the phenotype of each individual represents a specific subset $\Im_{n,\ell}$ to be valuated by means of equation (9), which is used as the fitness function of the genetic algorithm.

In order to enhance the effectiveness of the applied genetic algorithm, a specific local search operator, namely Digital Hill Climbing Operator (DHCO) (Papadakis 2002) is applied to the elite individual of each generation. This operator selects randomly a relatively small number of bits (i.e. 4 bits) from the elite's chromosome and generates all possible chromosomes (15 chromosomes) by permuting the selected bits and keeping the rest ones intact. The chromosome with the best fitness value is adopted as the new elite chromosome replacing the initial one.

4. Experimental results. In order to clarify some functional aspects of the proposed approach, three artificial examples are investigated at first. Then, the method is applied to two real world data sets. Finally, the proposed input selection method is applied to an industrial problem

4.1. Example I. Consider the following three-input, single output, non-linear system (Lin 1995):

$$y = [x_1^{1.5} - 1.5\sin(3x_2)]^2 + 5x_3 \mid x_1, x_2, x_3 \in [0, 3] \qquad (10)$$

Three hundred ($q = 300$) input/output data pairs $[x_1, x_2, x_3, y]^k$, $k = 1, 2, ..., 300$ were generated to create the initial data set $\wp_{q,m,1}$ that comprises $m = 3$ inputs and $q = 300$ data points. The first ($q/2 = 150$) data pairs were used to build the fuzzy surfaces for each input subset and the whole data set was used to valuate the importance of the respective input combination through equation (9). For each datum the input values are randomly generated within the interval [0,3] and the respective output is calculated according to (10). In (9), the value of the weight w equals $w = 0.8$. Since the number of candidate inputs is small, all the combinations were probed without genetic optimization. The simulation results are summarized in Table 1, ranked in ascending order with respect to their importance. The combination that comprises all inputs achieves the least value ($R_{3,1} = 9.10$) and hence, it is the most important one as explained in section 2.

Moreover, the valuation of the impact of each individual input to the output of the system is feasible by examining the simulation results for $n = 1$. For $n = 1$ the input x_2 achieves the smaller $R_{1,1}$ value among x_2 and x_3. The next most important single input is x_1 and follows input x_3. The order x_2, x_1, x_3 is also reported in (Lin 1995). The time required to evaluate all input permutations was 0.38 sec on a Pentium III 1.2Ghz computer.

Table 1. Input Evaluation for the system described by equation (10)

n	ℓ	$\Im_{n,\ell}$	$R_{n,\ell}$
3	1	x_1, x_2, x_3	9.10
2	1	x_1, x_2	25.04
2	2	x_2, x_3	28.13
1	1	x_2	46.51
2	3	x_1, x_3	47.65
1	2	x_1	65.80
1	3	x_3	80.06

Note that the results in Table 1 are not mutually exclusive. There is a reliability constraint introduced in (Leontaritis 1987) as follows: Let $\succ$ be a comparative operator regarding the relative importance of two inputs x_a, x_b. E.g. $x_a \succ x_b$ means that x_a is more important than x_b. The implication:

$$x_r x_t \succ x_s x_t \iff x_r \succ x_s \; \forall \; r, s, t \qquad (11)$$

has to be satisfied (Leontaritis 1987). i.e. $x_1 x_2 \succ x_1 x_3 \iff x_2 \succ x_3$. This reliability constraint is satisfied in Table 1.

It has to be stressed that implication (11) holds in the case where the inputs x_r, x_s, x_t are independent. In the case of interdependent inputs the implication (11) doesn't hold in general. The following interpretation is proposed. Let x_r, x_s, x_t be three inputs where x_r, x_t are interdependent. Also, let $x_r \succ x_s$. Since x_r, x_t are interdependent, the input x_t does not offer anything new to the combination x_r, x_t and, hence, x_r, x_t is nearly equivalent to x_r. On the contrary, input x_t jointly with x_s could be more important than x_r (or x_r, x_t), due to x_s, x_t independence.

4.2. Example II. The next example is a modification of the previous one. A dummy input is introduced to the input space, with no impact to the output of the system by modifying equation (10) as follows:

$$y = [x_1^{1.5} - 1.5\sin(3 \cdot x_2)]^2 + 5 \cdot x_3 \mid x_1, x_2, x_3, x_4 \in [0,3] \qquad (12)$$

The input data points were created by extending each input vector $[x_1, x_2, x_3]^k$, generated in the previous example by one more component that represents the x_4^k dummy-input value. Hence, 300 data points $[x_1, x_2, x_3, x_4, y]^k \mid k = 1, 2, ..., 300$ were produced as in Example I. After applying the proposed method, the best combination is x_1, x_2, x_3 with $R_{3,1} = 8.93$. The input variable x_4 is clearly rejected. The combination x_1, x_2, x_3, x_4 that includes the x_4 input was valuated with $R_{4,1} = 21.21$ which is 225% greater than the best combination. Additionally, when input x_4 was probed individually, it was valuated with $R_{1,4} = 85.43$ as the worst option.

In the following, the proposed method is compared to two other feature selection techniques from the literature, namely, the principal component analysis (PCA) (Haykin 1994), and the structure I identification method suggested in (Sugeno 1993). The PCA method, also known as the Karhunen-Loeve transformation (K-L transform) in communication systems, applies an invertible linear transformation T to the original input vector $\mathbf{x}$ such that the transformed vector $T\mathbf{x}$ provides maximal variance truncation with regard to the vector components, in the mean square sense. The PCA leads to dimensionality reduction of the input space, with the selection of the effective features decided by observing the size of the eigenevectors of the correlation matrix formulated by $\mathbf{x}$. Applying the PCA method, we are led to the input set $\underline{x} = [x_1, x_2, x_3, x_4]^T$. The resulting combination includes all input candidates, thus failing to recognize the redundant component x_4. As expected, this is a reasonable outcome, since PCA focuses on the correlation analysis of the input components, ignoring their relevance to the model's output. Finally, the input selection method suggested in (Sugeno 1993) belongs to the family of sequential forward search (SFS) techniques, proceeding along the following steps. At level 1, we search for a sole input component, being optimal among the set of candidate input set. At subsequent levels, the optimal set, currently available, is augmented by introducing an additional input component. The process is iteratively repeated to higher levels, until the optimal input

subset is finally attained. At each level of the process, evaluation of a specific input combination is carried out on the basis of the regularity criterion, requiring the time consuming construction (due to training process) of two separate fuzzy models. Applying the above method, the correct input combination x_1, x_2, x_3 was derived. Nevertheless, the above result was obtained at the expense of considerably larger computational cost when compared to our method. Additionally the input combination obtained, depends on the order the input variables were considered in the selection process, thus leading to suboptimal solutions.

4.3. Example III. In this example we examine the case of interdependent inputs. Suppose we have a system governed by the non linear equation:

$$y = \sin(x_1) + \sin(x_2) + 10^{-4} \cdot x_3 \ \text{ where } x_2 = 3 \cdot \sin(x_1) \qquad (13)$$

Inputs x_1, x_2 are interdependent, by means of equation $x_2 = 3 \cdot \sin(x_1)$ (or $x_1 = \sin^{-1}(x_2/3)$) while the input x_3 affects slightly the output. Three-hundred input-output data pairs were generated in a manner similar to Example I. The input vector $[x_1, x_2, x_3]^k$, $k = 1, ..., 300$ comprises random numbers within the range [0,3] and the corresponding output y^k is calculated by equation (13). The evaluation of the inputs indicates x_2 as the best option with $R_{1,1} = 2.37$. The combination x_1, x_2 was valuated with $R_{2,1} = 2.43$. Hence, both x_1 and x_3 inputs were rejected.

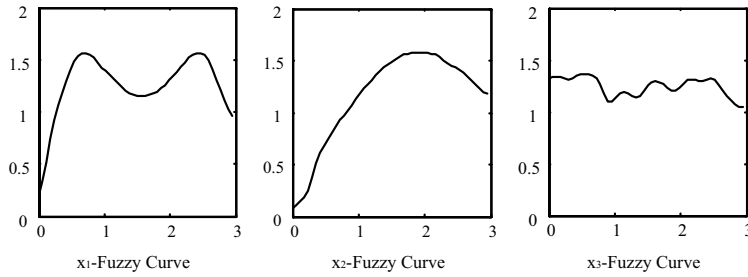x₁-Fuzzy Curve   x₂-Fuzzy Curve   x₃-Fuzzy Curve

Fig. 2. The Fuzzy Curves of variables x_1, x_2, x_3. Both x_1, x_2 inputs are considered significant.

We also applied the PCA method in this example. According to the eigenvalue ordering we were lead to the input ranking x_3, x_2, x_1. Hence, PCA suggests that x_3 is the most significant input because it corresponds to the largest eigenvalue.

Owing to the inherent drawback of the PCA approach, the method was unable to recognize the small impact of x_3, as well as the interdependency of inputs x_1, x_2. Finally, the method suggested in (Lin 1995) is applied, where the decision as to which inputs should be selected is made by means of the fuzzy curves associated to each input component. The method rejected input x_3 since a fuzzy curve with negligible height is observed due to the small exponent multiplying

this input in equation (13). Nevertheless, the method suggests that both inputs x_1 and x_2 should be regarded as significant, since both variables provide fuzzy curves with sizable heights as shown in Fig. 2. Although the approach presented in (Lin 1995) exhibits a linear time complexity with respect to the number of inputs, it is inefficient because the input candidates are independently evaluated, thus failing to detect a potential interdependence between several inputs.

4.4. The Fisher Iris benchmark. This benchmark data set comprises four candidate inputs that represent measured attributes of a crinum family such as x_1:sepal-length, x_2:sepal-width, x_3:petal-length, x_4:petal-width. The lilies are classified into three categories (Iris sestosa-versicolor-virginica) represented by an integer number from 1 to 3. The data set consists of 150 input-output data points. Our method is applied to track down the attributes, which are necessary and sufficient to describe a classification system. The first 75 points were used for building the fuzzy surfaces. The proposed input selection method suggests the combination x_1, x_2, x_3, x_4 as the best one, with $R_{4,1} = 3.10$. The next best in order was x_3, x_4 with $R_{3,1} = 3.18$. All combinations that include the x_2 input obtain relatively large values, while the x_2 input was the worst option when valuated alone. Following the methodology presented in (Lin 1995), the input x_2 would be rejected due to a flat fuzzy curve although its combination with the other three inputs affects significantly the output of the system. The latter is a major advantage of the algorithm proposed here.

Introducing a pseudo input to the set of candidates inputs as in Example II, and reapplying our input selection algorithm the combination x_1, x_2, x_3, x_4 is recognized again as the most important one acquiring a valuation $R_{4,1} = 2.46$. The combination x_3, x_4 is again the next best option with $R_{2,1} = 2.52$, while the combination x_1, x_2, x_3, x_4, x_5 that includes the pseudo input x_5 is worse enough, obtaining an evaluation measure value $R_{5,1} = 13.85$. Moreover, the input x_5, when probed alone, is ranked in the last place.

4.5. The gas furnace problem. The Box & Jenkins (Box 1970) gas furnace problem, a well-known real world application, has been considered. This dynamic problem involves a single input $u(t)$ that represents the gas flow at time t, and a single output $y(t)$ which corresponds to the CO_2 concentration in the exhaust gas of the furnace. The objective of many modeling techniques (Box 1970, Barada 1998, Lin 1995, Sugeno 1993) has been the prediction of $y(t)$, using past values of both the input $u(t)$ and the output $y(t)$.

To this end, a fuzzy model is sought in the functional form described by equation (5), providing one-step ahead predictions of the output $y(t)$. The model inputs are past values of the external input and the output, as shown in Fig. 1. The philosophy underlying the formulation of the input vector is described as follows: Since the proposed fuzzy model (equation (3)) is non-linear but static in principle, there is a need to formulate the dynamic nature of the process under consideration, in an efficient manner. Embedding the process dynamics is accomplished

Table 2.   Comparison of the proposed input selection approach

Method	Input Selection	R	Training Error (mse)	Checking Error (mse)
Our Method	$u(t-2), u(t-4), y(t-1)$	26.83	0.032	0.68
(Sugeno 1993)	$u(t-3), u(t-4), y(t-1)$	28.61	0.031	0.73
(Lin 1995)	$u(t-5), u(t-6), y(t-1)$	42.77	0.070	1.12
(Barada 1998)	$u(t-4), u(t-5), y(t-1)$	35.60	0.061	0.67

through the presentation to the model of selected projections of the time varying events onto a spatial representation of parallel inputs. The above approach corresponds to discretizing the process signals, considering past time windows of adequate size, such that a significant amount of information is preserved. Actually, a transformation is performed from the time domain to the spatial domain through the parallel presentation of the inputs. Hence, despite the fact that the model is static, dynamic systems can be effectively handled, provided an appropriate transformation. In other words, an input selection mechanism should be developed, selecting the input and output values at proper time instants in the past. These values are then handled by the model, providing future estimates of the system's output. Note that the past time horizon of the model may have a significant impact on the complexity of the resulting fuzzy model.

In the present case, the candidates input set consists of 20 inputs $u(t), u(t-1), ..., u(t-10), y(t-1), y(t-2), ..., y(t-10)$ that is $2^{20} - 1 = 1,048,575$ input combinations. The targeted output is $y(t)$. The genetic algorithm presented in section 3 is applied to locate the combination with the minimum R-value. The chromosome of the algorithm comprises 20 bits and the fitness function is given by equation (9). A population of 25 individuals was employed. After 50 generations the best individual achieved a fitness score of $R_{3,1} = 26.83$ suggesting $\underline{x} = [u(t-2), u(t-4), y(t-1)]$ as the most effective input combination. It has to be pointed out that different methods in literature agree on the number of inputs to be used. However they propose a different subset of $n = 3$ inputs. In (Sugeno 1993), the proposed input subset was: $\underline{x} = [u(t-3), u(t-4), y(t-1)]$, in (Lin 1995) the three most important inputs were: $\underline{x} = [u(t-5), u(t-6), y(t-1)]$ and in (Barada 1998): $\underline{x} = [u(t-4), u(t-5), y(t-1)]$. In order to assess the validity of each result a feed forward neural network consisting of one hidden layer with 20 neurons, 3 inputs and 1 output is used for evaluating the different input combinations in a relatively fair manner, overcoming any complex modeling details related to each one. The Leverbeng-Marquat method is applied to train each neural network for 1000 epochs. The mean square error (mse) is used as a performance index to estimate both the identification and prediction quality of each network.

The effectiveness of our proposed input selection method is shown comparatively in Table 2, both in terms of the identification and also in terms of the gen-

eralization capabilities of the final model that will be used to model the process (in this case the neural network). Moreover, the coherence between the value of R and the effectiveness of the respective input combination has been confirmed. Finally, the proposed method exhibits low time complexity since the total simulation time was $4.33\ min$ on a Pentium III computer at 1.2Ghz. Moreover the total number of probed input combinations was: 50 generations $\times$ (25 evaluations per generation for reproduction+16 evaluations per generation for DHCO operator)=2048, which is far less than the total number of $2^{20} - 1$ input combinations.

4.6. An industrial modeling application. In this section the proposed method is applied to data prepossessing in the modeling of the pan granulator device in the Phosphoric Fertilizers Industry (PFI), Nea Karvali, Greece. The industrial problem is described in the following.

4.6.1. The industrial problem. The industrial production of nitrogenous fertilizers consists of two processes realized sequentially, namely "Wet Process" and "Dry Process". The former process produces an Ammonium Nitrate (AN) solution, which is fed to the latter one. More specifically, a highly concentrated hot AN melt is sprayed to the rotating pan granulator from a spraying nozzle manifold. The fertilizer "end product" consists of small fertilizer granules, each having size in the range of a few millimeters. The average size of the granules is the main factor that defines the fertilizer quality. A desired quality size can be obtained by tuning the values of several pan granulator operating parameters/variables and the latter are summarized in Table 3.

Work is currently in progress for developing a feedback automatic control mechanism for the pan granulator. In this context a dependable open loop model of the pan granulator is needed. It turns out that not all pan granulator operating parameters in Table 3 are important as explained below.

4.6.2. Data acquisition. Data samples have been collected during the last five years for several fertilizer types. More specifically, several pan granulator operating variables have been sampled manually every two hours around the clock. In addition, the corresponding average (fertilizer granule) diameter size has been recorded. All the data corresponded to a steady state operation of the pan granulator.

The data used in this work included samples of fourteen operating variables involved in the production of fertilizer type CaN26 during late April/early May 2003 in the Phosphoric Fertilizers Industry (PFI). The aforementioned variables are shown in Table 3, including the corresponding units; in particular note that the units for the Nozzle Vertical Distance (rings) and the Spraying Angle (lines) are customized units used in the industry. A number of 174 data vectors had been available. Twenty data vectors including one (or more) missing values have been removed in the context of this work.

Table 3. Operating variables available for modeling the operation of the pan granulator in the Phosphoric Fertilizers Industry (PFI).

#	Variable name	Unit	Selected Features
1	AN Melt Flow	m^3/h	$\surd$
2	Recycled Fertilizer	T/h	$\surd$
3	AN Melt Temperature	^{o}C	$\surd$
4	AN Melt Pressure	bar	$\surd$
5	Granulation Temperature	^{o}C	−
6	Pan Inclination	degrees	−
7	Pan Rotation Speed	Hz	$\surd$
8	Nozzle Vertical Distance	rings	−
9	Nozzle Distance from the pan	cm	$\surd$
10	Scraper Speed	Hz	−
11	Spraying Angle	lines	−
12	Coarse Screen Vibration	%	−
13	Fine Screen Vibration	%	−
14	$Mg(NO_3)_2$ Supply	%	−

4.6.3. Input selection. Using the method presented here eight of the operating variables shown in Table 3 have been discarded. The selected six operating variables are marked by ($\surd$) in Table 3. That is a significant reduction of complexity from a 14-dimensional input space to a 6-dimensional input space. The practical significance of the selected variables has been confirmed by interviewing human operator experts in the industry. It might be interesting to point out that a visual inspection of the operating variables samples revealed that the samples of both the Coarse- and the Fine- Screen Vibration variables have all been constant equal to 80%, hence our variable selection method discarded the latter variables right. Due to the scope of this paper further modeling details will be presented in future work.

5. Conclusion. This work has presented a novel, fast and consistent (in terms of equation (11)) methodology for calculating the importance of a subset of inputs, from a set of candidate inputs, which jointly influence the output of a system. The proposed methodology employs the fuzzy surface technique, that is an enhanced extension of the fuzzy curve technique from the literature (Lin 1995). A genetic search has been used for probing combinations of inputs. Extensive simulation results, on both benchmark and real world data, have demonstrated the effectiveness of the proposed methodology. Future work will detail industrial fuzzy system modeling applications, where input variables will be selected using the methodology proposed here.

REFERENCES

Barada S., and Sing H. (1998). *Generating optimal adaptive fuzzy-neural models of dynamical systems with application to control, IEEE Trans on Syst. Man and Cyb.*, Part c, vol 28, no 3, pp. 371-391.

Battiti R. (1994). *Using mutual information for selecting features in supervised neural net learning, IEEE Trans. on Neural Networks*, vol. 5, no. 4, pp. 537-550.

Box G. E. P., and Jenkins G. M. (1970). *Time series analysis, forecasting and control*, San Francisco, CA: Holden Day.

Bradley P.S., Mangasarian O.L., Street W.N., (1998). *Feature selection in mathematical programming, INFORMS Journal on Computing* vol. 10, no. 2, pp. 209-217.

Chen S., and Billings S.A. (1989). *Representation of non-linear systems: The NARMAX model, Int. J. control*, vol. 49, no. 3, pp. 1013-1032.

Feng Chia Juang. (2002). *A TSK - Type Recurrent Fuzzy Network for Dynamic Systems Processing by Neural Network and Genetic Algorithms, IEEE Trans. on Fuzzy Syst.*, vol. 10, no. 2, pp. 155-170.

Fukunaga K. (1990). *Introduction to Statistical Pattern Recognition*, second ed., Academic Press, Boston.

Haykin Simon. (1994). *Neural Networks. A Comprehensive Foundation*, Macmillan Colledge Publishing Company, Inc.

Hopkins C., Routen T., Watson T. (1994). *Problems with using genetic algorithms for neural network feature selection,* 11th European Conference on Artificial Intelligence. pp. 221-225.

Kittler J. (1975). *Mathematical methods of feature selection in pattern recognition, International Journal of Man-Machine Studies* vol. 7, pp. 609-637.

Lee C. C. (1990). *Fuzzy Logic in Control Systems: Fuzzy Logic Controller- Part I, IEEE Trans. on Syst. Man and Cybern.*, vol. 20, no. 2, pp. 212-241.

Leontaritis I.J., and Billings S.A. (1985). *Input output parametric models for non linear systems, Int. J. Control*, vol 41, no 5, pp. 303-344.

Leontaritis I.J., and Billings S.A. (1987). *Model selection and validation methods for non-linearsystems, Int. J. control*, vol 45, no 1, pp. 311-341.

Lin C.T., and Lee C.S.G. (1992). *Real-Time Supervised Structure/Parameter Learning for Fuzzy Neural Network,* IEEE Conf. on Fuzzy Systems, pp. 1283-1291.

Lin Y., and Cunningham G. A. (1995). *A New Approach to Fuzzy-Neural System Modeling, IEEE Trans. Fuzzy Syst.*, vol. 3, pp. 190-198.

Malki H.A., Moghaddamjoo A. (1991). *Using the Karhunen-Loeve transformation in the back-propagation training algorithm, IEEE Trans. on Neural Networks*, vol. 2, no. 1, pp. 162-165.

Papadakis S., and Theocharis J. (1996). *A Novel Genetic Training Algorithm With Application to Fuzzy Neural Networks,* IASTED Int. Conf. On Syst. Modeling Identification and Control pp. 125-128.

Papadakis S. E., and Theocharis J. B. (2002). *A GA based Fuzzy Modeling Approach for Generating TSK models, Fuzzy Sets and Systems*, vol. 131, no. 2, pp. 121-152.

Piramuthu S. (2004). *Evaluating feature selection methods for learning in data mining applications, European Journal of Operational Research*, vol. 156, pp. 483-494.

Siedlecki W., Sklansky J. (1998). *A note on genetic algorithms for large-scale feature selection, Pattern Recognition Letters*, vol. 10, no. 5, pp. 335-347.

Sugeno M., and Yasukawa T. (1993). *A fuzzy-logic-based approach to qualitative modeling, IEEE Trans. Fuzzy Syst.*, vol. 1, pp. 7-31.

Wang L.X. (1992a). *Fuzzy systems are universal approximators,* in Proc. IEEE Int. Conf. Fuzzy Systems (San Diego, CA), pp. 1163-1170.

Wang L. X., and Mendel J. M. (1992b). *Back-propagation fuzzy system as non-linear dynamic system identifiers*, IEEE Conf. on Fuzzy Systems, pp. 1409-1416.

Wang L., and Langari R. (1995). *Building Sugeno-type models using fuzzy descretization and orthogonal parameter estimation techniques, IEEE Trans. Fuzzy Systems*, vol. 3, no. 4, pp. 454-458.

Stelios Papadakis received the Dipl. Eng. degree from the Department of Electrical Engineering at the Aristotle University of Thessaloniki, Greece in 1994 and the Ph.D. degree in 1999. His research interests include combinatorial optimization techniques, Fuzzy inference systems, neural networks, and the modeling of non-linear dynamical industrial systems for prediction and control.

Panagiotis Tzionas received the B.Eng Degree in Electrical/Electronic Engineering from Imperial College, University of London, the M.Sc Degree in Digital Systems form King's College, University of London and the Ph.D degree from the Dept. of Electrical and Computer Engineering, University of Thrace, Greece. Currently he serves as a Professor at the Dept. of Automation, Technological Educational Institute of Thessaloniki. His research interests are mainly in the area of neural networks and fuzzy systems and their implementations.

Vassilis G. Kaburlasos received the Diploma degree from the National Technical University of Athens, Greece, in 1986, and the M.Sc. and Ph.D. degrees from the University of Nevada, Reno, in 1989 and 1992, respectively, all in electrical engineering. He has been participant investigator or leader in 15 research projects, funded both publicly and privately, in the U.S.A. and in the European Union. He has authored or coauthored more than 55 scientific research papers. He is currently an Associate Professor. His research interests include intelligent machine modeling, design and industrial applications. Dr. Kaburlasos is a member of ACM, INNS, the Technical Chamber of Greece, Sigma Xi, Phi Kappa Phi, Tau Beta Pi, Eta Kappa Nu.

John Theocharis graduated as an Electrical Engineer from the Aristotle University of Thessaloniki in 1980. From 1980 to 1985 he has been with the scientific staff of the Department of Electrical and Computer Engineering where he received the Ph.D. degree in 1985. Since 1990 he is an Associate Professor at the Department of Electronics and Computer Engineering at the same University. His research interests include adaptive control, intelligent systems and electrical motor drives. Recently, he has focused his activity to neural networks, fuzzy logic control, fuzzy neural networks with applications to identification and control of complex nonlinear systems.

A genetic based approach to the Type I structure identification problem

S.E. Papadakis, P. Tzionas, V.G. Kaburlasos, J.B. Theocharis

The problem of system input selection, is dealt with here using an effective novel method based on fuzzy surfaces. A simple genetic algorithm considers jointly more than one input, for finding an optimal subset of necessary and sufficient inputs. Extensive simulation results on both benchmark data and real world data have demonstrated comparatively the advantages of the proposed method.