Mathware & Soft Computing 5 (1998) 305-319

Learning from Imprecise Examples with GA-P Algorithms

Luciano Sánchez[1] & Inés Couso[2]
[1]Dept. de Informática & [2]Dept. de Estadística e I.O y D.M.
Universidad de Oviedo
e-mail: luciano@lsi.uniovi.es & couso@pinon.ccu.uniovi.es

Abstract

GA-P algorithms combine genetic programming and genetic algorithms to solve symbolic regression problems. In this work, we will learn a model by means of an interval GA-P procedure which can use precise or imprecise examples. This method provides us with an analytic expression that shows the dependence between input and output variables, using interval arithmetic. The method also provides us with interval estimations of the parameters on which this expression depends.

The algorithm that we propose has been tested in a practical problem related to electrical engineering. We will obtain an expression of the length of the low voltage electrical line in some spanish villages as a function of their area and their number of inhabitants. The obtained model is compared to statistical regression-based, neural network, fuzzy rule-based and genetic programming-based models.

1 Introduction

Learning a model from a set of precise examples $\{(X^1, Y^1), (X^2, Y^2), \ldots\}$ and a given family of models $\mathcal{G}$ can be described as finding the model $g \in \mathcal{G}$ for which the similarities between the values $g(X^i)$ and the desired responses Y^i are the highest ones. When a parametric family $\mathcal{G}_\theta = \{f_\theta\}_\theta$ is used, the learning process requires estimating the value of the parameter θ^* that makes $g = f_{\theta^*}$ the best choice, by mean of a suitable analytical or numerical procedure. Linear regression and some kinds of neural network training methods belong to this type of problems.

If the expression of the function that defines the model is not known previously and we need to determine both this expression (we will call it the "structure" of the model) and the best values of the parameters on which it depends, the technique is known as *symbolic regression.* Solving symbolic regression problems is one of the main concerns of genetic programming [6].

On the other hand, learning models from imprecise examples is quite a different problem. We will use intervals to represent imprecision (i.e. "The output is between

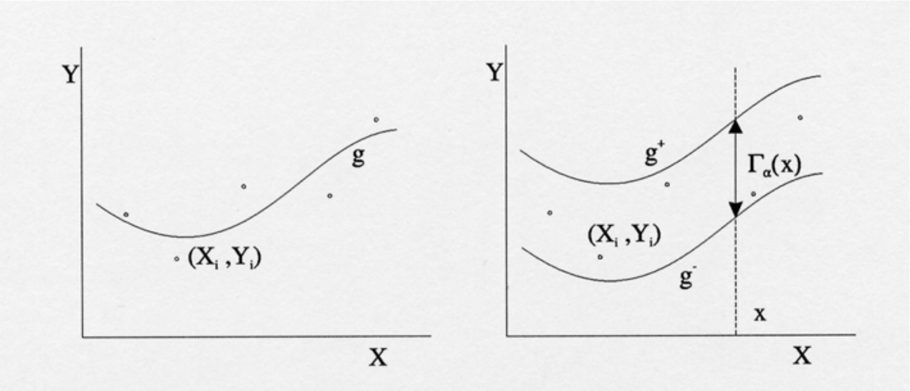

Figure 1: Punctual and intervalar models. Intervalar models produce a range of values for each input. That range contains the true value of the output with a high probability. An intervalar model is equivalent to two punctual models that give upper and lower limits of the range of outputs.

1 and 3") and *interval models* to catch the relationship between input and output data. The output of an interval model is a range of values delimited by two functions g^- and g^+ (see Figure 1). The output of the model is requiered to contain the desired responses, i.e. $g^-(X^i) \leq Y^i \leq g^+(X^i)$ with high probability, while keeping the amplitude of the outputs $g^+(X^i) - g^-(X^i)$ as low as possible.

There are some works related to intervalar neural networks, and we take into account that α-cuts of fuzzy models are also intervalar models [8] but, as far as we know, symbolic regression methods have only been used to find *punctual* estimates. We will extend these methods so that they can provide us with interval values.

2 Intervalar predictions

2.1 Multi-valued predictions

When learning a punctual model, we search for a function g such that the difference $Y - g(X)$ is small for every value of X; in other words a function g so that $g(X)$ is a good estimate of Y.

From a stochastic point of view we can assume that there exists a random experiment governed by a probability measure P with results in a set Ω such that $Y : \Omega \longrightarrow \mathbb{R}$ is a random variable and $X : \Omega \longrightarrow \mathbb{R}^d$ is a random vector. The function g that minimizes the mean square error in that case is $g(x) = E[Y \mid X = x]$ [9].

In some practical problems it is also interesting to obtain the margins in which we expect the variable Y is when the variable X (which can be multidimensional) is known. When we need to solve the punctual problem, we search for a function g such that $g(X)$ estimates $E[Y \mid X]$. Now we need an interval of values Γ_β that covers the value Y with probability higher than a confidence degree β and this interval must depend on the value of the variable X, so Γ_β is a function of X. Since X is also a function of ω, this mapping between the result ω of a random experiment and an interval $\Gamma_\beta(X(\omega))$ is a random set.

Formally, we will look for a multi-valued mapping $\Gamma_\beta : \mathrm{Im}(X) \longrightarrow I(\mathbb{R})$, where $I(\mathbb{R})$ is the set formed by all closed intervals in $\mathbb{R}$, such that the random set $\Gamma_\beta \circ X : \Omega \longrightarrow I(\mathbb{R})$ verifies

$$P\{\omega \in \Omega \mid Y(\omega) \in \Gamma_\beta \circ X(\omega)\} \geq \beta$$

for a given degree of confidence β (the symbol "$\circ$" means composition: $\Gamma_\beta \circ X(\omega) := \Gamma_\beta(X(\omega))$ and β is the probability that Y is in the interval $\Gamma_\beta(X)$.)

We can assess an interval prediction in some different ways. For example, we can say that given a value for β, the shorter Γ_β is, the better it is. Let us define two functions g^+ and g^- so that $g^-(X)$ is the minimum value of the confidence interval $\Gamma_\beta(X)$ and $g^-(X)$ is the maximum,

$$\Gamma_\beta \circ X = [g^- \circ X, g^+ \circ X]$$

and let us impose that g^+ and g^- are continuous (see Figure 1). Then, the margin of validity will be better when the mean difference between g^+ and g^-

$$E(g^+(X) - g^-(X))$$

is low. Since it must be true that

$$P\{\omega \in \Omega \mid g^-(X(\omega)) < Y(\omega) < g^+(X(\omega))\} \geq \beta$$

we can define the objective of the interval prediction as "find two functions g^+ and g^- such that the distance between $g^+(X)$ and $g^-(X)$ is minimum and Y is between $g^+(X)$ and $g^-(X)$ with a probability β".

In other words, given a region

$$R_{(g^+,g^-)} = \{(x,y) \in \mathbb{R}^{d+1} \mid g^-(x) < y < g^+(x)\}$$

we need to minimize

$$E(g^+(X) - g^-(X))$$

constrained to

$$P\{\omega \in \Omega \mid (X,Y)(\omega) \in R_{(g^+,g^-)}\} \geq \beta.$$

If we are solving an interval modelling problem we need to find two functions g^+ and g^- instead of the single function g that we needed to find in punctual modelling. Let us suppose now that we define g^+ and g^- by means of a function of X that depends on some interval parameters, using interval arithmetic [1]. This concept is similar to that introduced in [7] and many other works related to fuzzy regression [8]. Formally, let g^+ and g^- depend on a function $h_\theta : \mathbb{R}^m \longrightarrow \mathbb{R}$ so that $[g^-(x), g^+(x)] = \{t \in \mathbb{R} \mid t = h_\theta(x), \theta \in [\theta_1^-, \theta_1^+] \times \ldots \times [\theta_m^-, \theta_m^+]\}$ where the expression of h_θ is known except for the value of $2m$ parameters θ_k, $k = 1, \ldots 2m$ and h_θ is continuous with respect to θ and x (and then g^+ and g^- will also be continuous functions, as we had proposed). Given a function h, a random sample of size N obtained from the random vector (X, Y),

$$((X_1, Y_1), \ldots, (X_N, Y_N))$$

(where (X_i, Y_i) are independent and identically distributed) and a confidence degree $1 - \epsilon$ we can estimate $\theta_i^-(\epsilon)$ and $\theta_i^+(\epsilon)$ with the $2m$ values that minimize

$$\frac{1}{N}\sum_{i=1}^{N}(g^+(X_i) - g^-(X_i))$$

constrained by

$$1 - \epsilon \leq \frac{1}{N}\#\{i \in \{1 \ldots N\} \mid (X_i, Y_i) \in R_{(g^+, g^-)}\}$$

that is, the number of elements in the sample that belong to $R_{(g^+, g^-)}$.

For a given value of ϵ we can estimate the value of β by means of a second sample

$$((X_1', Y_1'), \ldots, (X_M', Y_M')),$$

independent from the first one, by means of

$$\hat{\beta}_M = \frac{1}{M}\#\{i \in \{1 \ldots M\} \mid (X_i', Y_i') \in R_{(g^+, g^-)}\}.$$

The random variable $M \cdot \hat{\beta}_M$ follows a binomial distribution with parameters M and β and, by the strong law of the large numbers, it converges almost surely to the value β when $M \to \infty$.

The practical implications of those definitions are straightforward: let us suppose that we define a family of interval valued models and we use a learning procedure to choose the model that best describes a set of N *precise* examples. We first choose an arbitrary value ϵ, and we decide that the output of any valid model must contain the desired output in more than $(1 - \epsilon)N$ examples. If we also decide that the sum of the amplitudes of the output of the model in those $(1 - \epsilon)N$ points defines how good the model is, we only need a procedure that searches over the space of valid models the minimum of the sum of these amplitudes. But we will obtain a model for which the expected fraction of covered points will be unknown and likely to be higher than $(1 - \epsilon)$. This fraction (which is a sort of "generalization error") can be estimated by the proportion of points that are covered by the output of the model when confronted to a new sample; in other words, the generalization error is estimated by one-fold cross validation.

The proposed procedure takes three steps:

1. Choose the familiy of functions $\{h_\theta\}_\theta$ and the value of ϵ.

2. Estimate the model (i.e., the values of the interval parameters $[\theta_i^-, \theta_i^+]$) from the first sample, which we will call "training set"

3. Estimate the confidence degree β from the second sample (test set)

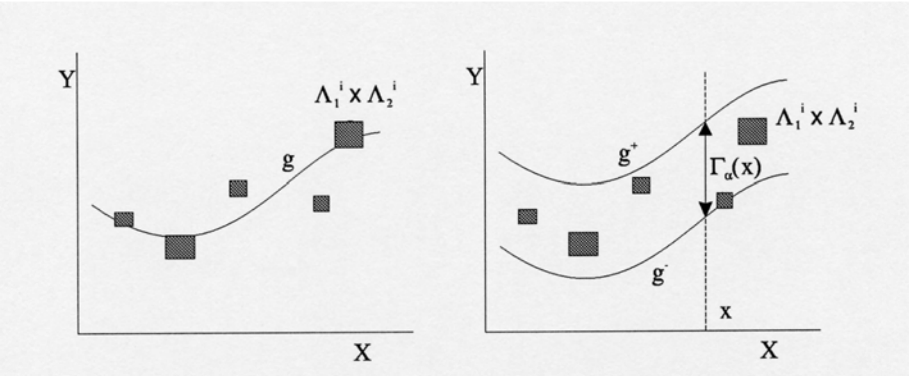
Y
$\Lambda_1^i \times \Lambda_2^i$
g
X
Y
g^+
$\Lambda_1^i \times \Lambda_2^i$
$\Gamma_\alpha(x)$
g^-
x
X

Figure 2: Punctual and intervalar estimation from imprecise data. When data are imprecise it is difficult to learn a punctual model, but we can still give a range of values than contain the output, given an imprecise input, with high probability.

2.2 Imprecisely measured data

Let us suppose that the values of the random variable Y and the random vector X cannot be precisely observed but we only know that for a given output ω of the random experiment

$$(X,Y)(\omega) \in \Lambda(\omega),$$

where $\Lambda = \Lambda_1 \times \Lambda_2$, with $\Lambda_1 : \Omega \longrightarrow I(\mathbb{R}^d)$ and $\Lambda_2 : \Omega \longrightarrow I(\mathbb{R})$ are random sets, and $I(\mathbb{R}^d)$ is the set of all the rectangles in $\mathbb{R}^d$ (see Figure 2.) For example, imagine that we have a sensor that indicates "between 100 and 110" when $X(\omega_1) = 100$ and also when $X(\omega_2) = 105$; we model this behaviour by means of a random set Λ_1 such that $\Lambda_1(\omega_1) = [100, 110]$ and $\Lambda_1(\omega_2) = [100, 110]$, and it is true that $X(\omega_1) \in \Lambda_1(\omega_1)$ and $X(\omega_2) \in \Lambda_1(\omega_2)$.

In these conditions, there is not an extension of the classical modeling that is universally accepted as the best one. But the previous model can deal with this imprecision. Let us define two functions g^+, g^- such that $P\{\omega \in \Omega \mid g^-(X(\omega)) < Y(\omega) < g^+(X(\omega)) \quad \forall (X,Y) \in C(\Lambda)\} \geq \beta$ and let $C(\Lambda) = \{U$ random variable $\mid U(\omega) \in \Lambda(\omega) \quad$ a.s. $(P)\}$ be the set of all random variables *contained* in Λ (see [2]). The set $C(\Lambda)$ includes all possible mappings (random variables) that can relate a result ω of the random experiment with values $X(\omega)$ and $Y(\omega)$ that are compatible with the imprecise observations Λ_1 and Λ_2.

We wish that the mean margin between g^+ and g^- is the narrowest possible one for a given β, but now it is posed an additional difficulty, because we do not know (X,Y) but a random set Λ which contains it. Taking a pessimistic criterion, we search for a region $R_{(g^+,g^-)} = \{(x,y) \in \mathbb{R}^{d+1} \mid g^-(x) < y < g^+(x)\}$ for which all points in the set $P_\Lambda(R_{(g^+,g^-)}) = \{P_{(X,Y)}(R_{(g^+,g^-)}) \mid (X,Y) \in C(\Lambda)\}$ $= \{t \in [0,1] \mid t = P\{\omega \in \Omega \mid g^- \circ X(\omega) < Y(\omega) < g^+ \circ X(\omega)\}, (X,Y) \in C(\Lambda)\}$ are higher than the confidence β. For every pair of variables (X,Y) contained in Λ we obtain a value for the probability that (X,Y) is in $R_{(g^+,g^-)}$: the set $P_\Lambda(R_{(g^+,g^-)})$

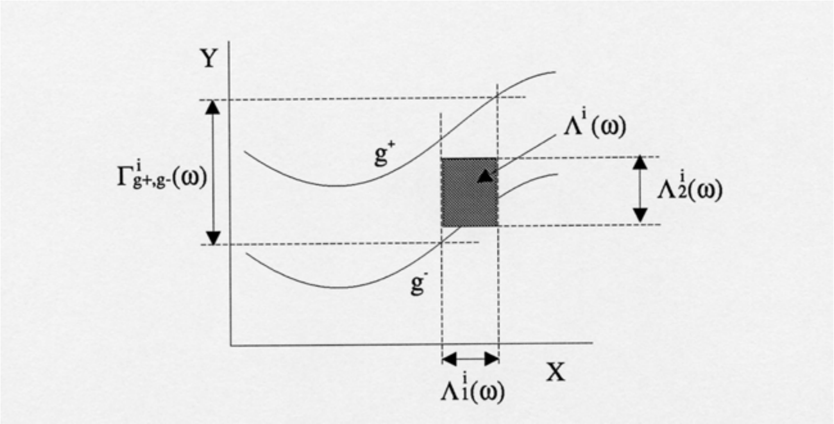

Figure 3: Calculations of the value of $\Gamma \circ \Lambda_X$. When the input value is imprecise, the output of an intervalar model is the projection over the output space of the intersection between the cylindrical extension of the input and the "graph" of the model. This mechanism is very similar to the used when making inference in fuzzy rule-based models.

is the set of all these values and it is bounded by the numbers β^- and β^+ [2] where

$$\beta^- = P\{\omega \in \Omega \mid \Lambda(\omega) \subset R_{(g^+,g^-)}\}$$

$$\beta^+ = P\{\omega \in \Omega \mid (\Lambda(\omega) \cap R_{(g^+,g^-)}) \neq \emptyset\}.$$

Making

$$\beta^- \leq \beta$$

so that

$$\inf_{(X,Y)\in C(\Lambda)} P\{\omega \in \Omega \mid g^- \circ X(\omega) < Y(\omega) < g^+ \circ X(\omega)\} \geq \beta$$

we obtain a model that fulfills that the probability that Y is in the interval prediction is higher that β in the worst case.

When data were precisely observed, we tried to minimize the expected length of the random interval $[g^-(X), g^+(X)]$ constrained by $P[\omega \in \Omega \mid g^-(X(\omega)) < Y(\omega) < g^+(X(\omega))] \geq \beta$. This time we want to find the minimum expected length of the random interval $\Gamma_{(g^+,g^-)}$ (see Figure 3)

$$\Gamma_{(g^+,g^-)}(\omega) = \{y \in \mathbb{R} \mid y \in [g^-(x), g^+(x)] \ \wedge \ x \in \Lambda_1(\omega)\}$$
$$= [\min_{x\in\Lambda_1(\omega)} g^-(x), \max_{x\in\Lambda_1(\omega)} g^+(x)]$$

(where the last assertion is true because g^- and g^+ are continuous functions and we know $\Gamma_{(g^+,g^-)}$ is strongly measurable by the same reason) restricted to

$$\beta \leq \min_{(X,Y)\in C(\Lambda)} P\{\omega \in \Omega \mid g^-(X(\omega)) < Y(\omega) < g^+(X(\omega))\}.$$

To solve the problem, we propose the following estimation: let

$$(\Lambda^1, \ldots, \Lambda^N) = (\Lambda_1^1 \times \Lambda_2^1, \ldots, \Lambda_1^N \times \Lambda_2^N)$$

be a size N random sample drawn from the random set $\Lambda = \Lambda_1 \times \Lambda_2$. For a given sample, we choose a value $\epsilon > 0$ as before and also a function h_θ known except for the values of m parameters, and we search for $2m$ constants θ_i^-, θ_i^+ so that $[g^-(x), g^+(x)] = \{t \in \mathrm{I\!R} \mid t = h_\theta(x), \theta \in [\theta_1^-, \theta_1^+] \times \ldots \times [\theta_m^-, \theta_m^+]\}$ and the value $\widehat{L}_h = \frac{1}{N}\sum_{i=1}^N ||\Gamma^i_{(g^+,g^-)}||$ is minimum, where $\Gamma^i_{(g^+,g^-)} = \{y \in \mathrm{I\!R} \mid y \in [g^-(x), g^+(x)] \wedge x \in \Lambda_1^i\}$ and restricting the search to the set of functions g^+, g^- that fulfill $1 - \epsilon \leq \frac{1}{N}\#\{i \in \{1 \ldots N\} \mid \Lambda^i \subset R_{(g^+,g^-)}\}$. That is, the number of intervals in the sample that are contained in $R_{(g^+,g^-)}$.

Once g^+ and g^- have been found, we cannot estimate β but a range of values for β if we have a second independent sample

$$(\Lambda'^1_1 \times \Lambda'^1_2, \ldots, \Lambda'^M_1 \times \Lambda'^M_2)$$

for which the value

$$\hat{\beta}_M^- = \frac{1}{M}\#\left\{i \in \{1 \ldots M\} \mid \Lambda'^i \subset R_{(g^+,g^-)}\right\}$$

is an estimator of the *belief measure* [13] of the event "the imprecisely observed pair (X, Y) is in R", and

$$\hat{\beta}_M^+ = \frac{1}{M}\#\left\{i \in \{1 \ldots M\} \mid \Lambda'^i \cap R_{(g^+,g^-)} \neq \emptyset\right\}.$$

is an estimation of the *plausibility* [13] of the same event. In other words,

$$\hat{\beta}_M^- \xrightarrow[M\to\infty]{a.s.} P[\omega \in \Omega \mid \Lambda(\omega) \subseteq R_{(g^+,g^-)}] =$$

$$\inf_{(X,Y)\in C(\Lambda)} P_{(X,Y)}(R_{(g^+,g^-)})$$

$$\hat{\beta}_M^+ \xrightarrow[M\to\infty]{a.s.} P[\omega \in \Omega \mid (\Lambda(\omega) \cap R_{(g^+,g^-)}) \neq \emptyset] =$$

$$\sup_{(X,Y)\in C(\Lambda)} P_{(X,Y)}(R_{(g^+,g^-)}).$$

Finally, note that the case analyzed in the previous section is a particular case of this one (where $\Lambda_1 = X$ and $\Lambda_2 = Y$). Conversely, if a family of interval valued models is defined and a learning procedure used to choose the model that best describes a set of N *imprecise* examples, we must first choose an arbitrary value ϵ, and decide that the output of any valid model must *completely* contain the desired, imprecisely measured output in more than $(1 - \epsilon)N$ examples. The fitness of the model is similar to the mentioned in the preceeding section. Again, we will obtain a model for which the expected fraction of covered points will be higher than $(1-\epsilon)$ and this fraction can be pessimistically estimated by the proportion of examples in a new random sample that are contained in the graph of the model.

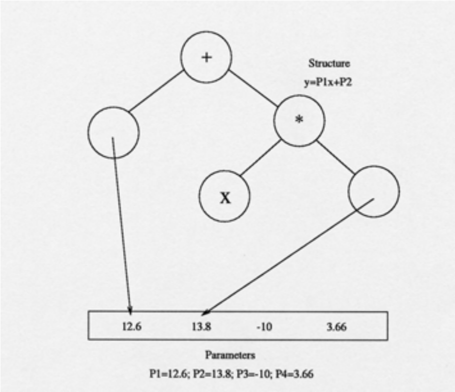

Figure 4: Individual in GA-P algorithms. It comprises two parts. The "structure" is defined by a tree that combines input variables, parameter names and algebraic operators. The values of the parameters are stored in a chain, like genetic algorithms do. The GA-P algorithm searches both a suitable structure and the values of the parameters that define the best model.

3 Numerical optimization method. GA-P Algorithms

The multivalued mapping Γ_{g^+,g^-} proposed in last section depends on a function h (whose expression is unknown to us), on $2m$ parameters θ_i^-, θ_i^+ and on the value $\hat{L}_h$ (which in turn depends on these parameters) and, by last, on a random sample drawn from the random set Λ. For a given random sample and a structure of the model (i.e., parametic expression for h), we can minimize $\hat{L}_h$ with respect to θ_i^- and θ_i^+, restricted to the conditions about the fraction of covered examples that were imposed before. The objective is to find the structure of the model with produces the lowest value of $\hat{L}_h$. But this problem is a particular case of symbolic regression, and we can apply genetic programming algorithms to solve it.

GA-P algorithms are an evolutionary computation method, hybrid between genetic algorithms and genetic programming, optimized to perform symbolic regressions. In this algorithm we begin with a population formed by set of possible solutions of the regression problem and by means of different operations, we create new solutions and discard others until one of them is good enough. A complete description of the GA-P method can be found in [4]. Briefly, we will remark that each element of the population comprises a chain of parameters (so called GA part, or coefficient part) and the description of a function (GP part, or expressional part) which depends on the parameters that are codified in GA part and that codifies the structure of the model.

The set of parameters contained in GA part is codified by means of any GA method (for instance, by means of an array of bits partitioned in sections, where every section represents a real value). The expressional part is a mathematical formula represented by a tree. In this component, internal nodes are mathematical operators (like +,−,×,/); when we evaluate the solution codified in an individual, terminal nodes are replaced by the value of its corresponding part in the chain of parameters or by the value of one of the input variables (see Figure 4).

The two basic operations by means of which new members of the population are generated are crossover and mutation. These operations are independently performed over GP and GA parts of an individual, and we use the same operators that GP and GA algorithms define.

3.1 Modifications to GA-P

The model proposed in last section will be defined in terms of a function h_θ and m interval parameters $[\theta_i^-, \theta_i^+]$. We need that the terminal nodes of the tree that codifies the expressional part can be intervals and we also need to program interval arithmetic operators to evaluate the expressional part of the GA-P; apart from this, the modifications to the GA-P codification scheme are straightforward.

The fitness function is not the same as the one used in conventional symbolic regression problems. It does not depend on the evaluations of the expressional part in the set of examples but on the mean separation between g_β^+ and g_β^-, as shown in the following expression:

$$\begin{cases} f_1 = K \cdot \sum [M(Y^i) - M(Y(X^i))]^2 & \text{if } \mathrm{N}_c < N(1-\epsilon_1) \\ f_2 = \sum ||Y(X^i)|| & \text{if } \mathrm{N}_c > N(1-\epsilon_2) \\ f_3 = p(N_c) f_1 + (1 - p(N_c)) f_2 & \text{else} \end{cases}$$

where $M([a,b]) = (a+b)/2$, $Y(X_i)$ is the output of the model for an input X_i, which can be expressed as

$$Y(X_i)(y) = \mathrm{proy}_{x \in \mathbb{R}^d}(R(x,y) \cap X_i(x)),$$

N_c is the number of samples contained in the graph R of the model:

$$\mathrm{N}_c = \#\{i \in \{1 \ldots N\} \mid Y_i \subseteq Y(X_i)\}),$$

K is a real value high enough so that $K \cdot \sum [M(Y^i) - M(Y(X^i))]^2$ is always higher than $\sum ||Y(X^i)||$ and $p(N_c) = (N(1-\epsilon_2) - N_c)/(N(\epsilon_1 - \epsilon_2))$.

The explanation of this function follows: when we are in the initial stages of the evolution, we replace each interval by its midpoint and search for the classical least squares solution. As soon as $N(1-\epsilon_1)$ examples are covered, we begin to promote those individuals that get a narrower band of prediction values. This fitness function always makes a model that covers more than $N(1-\epsilon_2)$ examples better than a model that covers less than $N(1-\epsilon_1)$, so population evolves gradually towards models with adequate covering that will be prefered on the basis of their mean amplitude.

We penalize the solutions that do not cover $N(1-\epsilon_2)$ examples by multiplying their fitness by a value K that is determined empirically. We think that it is not necessary to resort to multicriteria optimization (see [3]) because the value of K is not difficult to obtain. When the number of uncovered examples is lower than $N\epsilon_1$, the fitness is the mean amplitud of the model over the sample. In intermediate situations we promediate both values.

4 Practical application

The practical problem which inspired the method developed here follows. The problem will be solved by some different methods, and the obtained solutions will then be compared.

4.1 Introduction

In this work we will study the length of low voltage electrical line in rural villages, mainly located in mountain areas. These villages have a small number of inhabitants and have a reduced consumption of energy, so they have only one transformation center, with one, two or three main lines that serve all clients. The houses are very disgregated (there is not an "urban center" and the density of houses does not depend on the distance to the transformation center) and the ratio length of line – number of inhabitants is much higher than the measured in bigger villages or cities. Maintenance of the lines is more expensive too. Since this clients do not produce benefits, the companies that serve them are compensated with an amount that depends on the length of line installed. It is remarked than the company that paid for this study serves more than 10,000 villages of this type, so this measurement is relevant.

To validate the model of the line length in a village the company provided us with data: one file with the measured line length, the number of inhabitants and the mean of the distances from the transformation center to the three furthest clients in a sample of 491 rural nuclei. Our objective was to relate the line length with the other, firstly by classical methods and later by applying GA-P methods. A simple intervalar model that relates lengths with pairs clients-radius will be provided, and its performance will be compared to fuzzy rule based methods.

Our variables will be named as follows:

Symbol	Meaning
A_i	Number of clients in village i
R_i	Radius of village i
n	Number of villages in the sample
l_i	Line length, village i
$\tilde{l}_i$	Estimation of l_i
s_i	Number of sectors in village i

4.2 Application of classical methods

In order to apply classical methods, we needed to make some hypothesis. In the villages that are being studied, electrical networks are star-shaped and arranged in

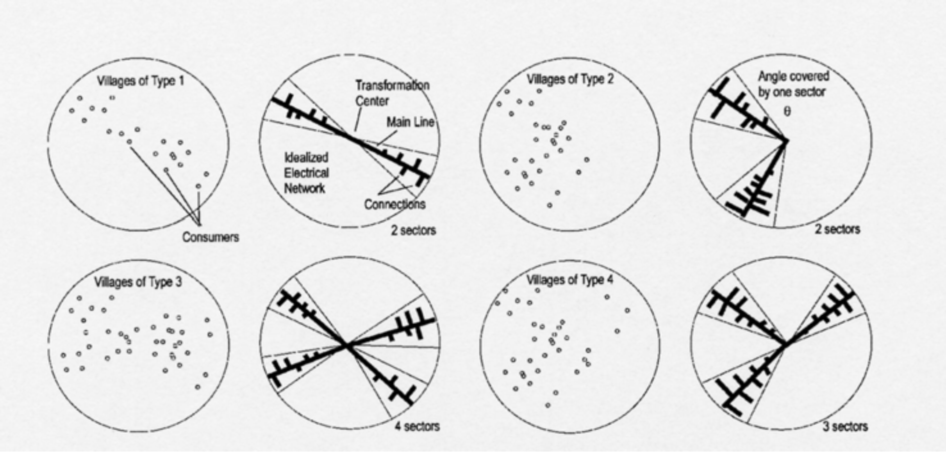

Figure 5: Idealized distributions of electrical networks

sectors. A main line passes near all clients inside them, and clients are connected to these main lines by small segments (see figure 5).

To build a theoretical simplified model we have admitted that:

- A village comprises s_i sectors. Each sector covers an angle $2\theta_i$. All sectors in the same village cover the same angle. Each sector is served by one output of the only transformation center in the village.
- All sectors in a village have the same radius, R_i.
- The density of clients is constant inside every sector.
- Inside a sector, the electrical line comprises a main line of length R_i and so many branches as consumers.

If we admit that customers are uniformly distributed, we can approximate the total length by multiplying the mean distance between one of them and the line by the number of inhabitants. Let us name this mean distance d_i for village i, and let the sector be $2\theta_i$ wide. Then $d_i = \frac{2(1-cos\theta_i)}{3\theta_i}R_i$ so cable length will be $\tilde{l}_i = s_i(R_i + \frac{A_i}{s_i}d_i) = s_iR_i + A_i\frac{2(1-cos\theta_i)}{3\theta_i}R_i$.

4.3 Classical regression

If the angles θ_i and the numbers s_i were similar enough between them, we could regard them as constants and estimate them by the parameters $\overline{\theta}_i = \theta$ and $\overline{s}_i = s$ of a least squares linear regression

$$\tilde{l}_i/R_i = s + k(\theta)A_i$$

to a set of pairs $(x, y) = (A_i, l_i/R_i)$.

Table 1: Cross-comparison of results

Method	Training	Test	Complexity
Linear	365	443	7 n., 2 p.
Linear, 2 classes	338	458	17 n., 6 p.
Exponential	342	426	7 n., 2 p.
2th order poly.	332	393	22 n., 6 p.
3rd order poly.	318	941	53 n., 10 p.
Interval GA-P	332	420	8 n., 2 p.
MLP 2-25-1	312	391	102 p.
W. M. fuzzy model	262	610	22 r.
TSK fuzzy model	272	462	34 r.
RSB fuzzy model [12]	241	410	39 r.

We can get a better fit by allowing a certain dependence between the number of sectors, their angles and the number of inhabitants. This can be done by dividing the sample into classes or by mean of a change of variables. Both cases were studied, and the best fit was obtained with the model

$$\frac{\tilde{l}_i}{R_i} = k_1 A_i^{k_2}$$

4.4 Interval GA-P

Let us apply GA-P algorithms to check whether we can obtain a formula that is comparable in complexity with the last one, while getting better fit to the real data. We will define "simple expression" as a formula that can be codified in a tree with no more than 20 nodes and depending on no more than 10 parameters. Binary operations will be sum, difference, product, ratio and power. The unary operation will be the square root.

We use the steady state approach, with tournament selection and elitism. The probability of crossover is 0.9, both in GP and GA parts. We do not perform mutation in the GP part and we apply this operator with probability 0.01 in GA part, which is encoded in floating point. We use local optimization (Nelder and Mead's simplex; this approach has not been used, as far as we know, in GA-P field, but there are some related works. See, for example [5]. Its original definition in GA field can be seen on [10, 11]) and overselection (1000 individuals). The population has fixed size, 100 individuals.

5 Comparison between methods

To compare classical, neural, fuzzy and GA-P methods we have divided the sample into two sets comprising 246 and 245 samples. If we had chosen a higher percentage of training data the numerical fit would have been better, but we want to compare

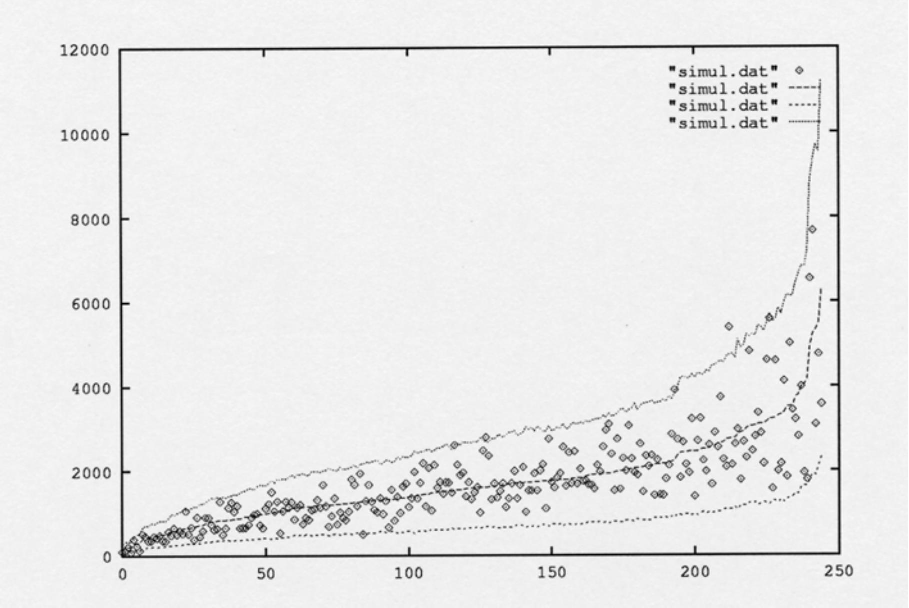

Figure 6: Output of the interval GA-P model and test set (squares). The line in the middle is the output of a punctual model built when replacing interval parameters by their midpoints. The x-axis show the ordinal number of the example. Examples are sorted by the value of the lower bound given by the model.

how well different methods generalize from few examples. In table 1 the mean square error values over these two sets are labeled *training* and *test*. The initials in the third row mean "nodes" and "parameters"; they are the number of nodes and parameters that we would need to codify the model like GP does.

Fuzzy rule based models were automatically designed by means of a genetic learning procedure. Observe that Wang and Mendel method and TSK cannot generalize from the example set that was chosen; their number of parameters is too high in relation to the size of the sample. Notice also that the TSK method is worse than the modelled labelled "linear", and that it is apparently incorrect, since the linear model is a particular case of TSK model. This is not an error; linear models shown here involve a change of variable and have one input variable but all the remaining methods were confronted with raw data. Observe also the inestability of high order polynomials and the good fit that the three layer perceptron gets.

By last, the model labeled "Interval GA-P" is as follows:

$$\begin{aligned}&Y([R^-,R^+],[A^-,A^+]) =\\ &[13.56,\ 53.47]\otimes[R^-,R^+]\otimes[0.058,0.064]^{[A^-,A^+]^{0.0625}}.\end{aligned}$$

Since this model cannot be directly compared to punctual ones, we evaluated the mean squared error of the punctual model

$$Y(R,A) = 30.23\, R\, 0.061^{(A^{0.0625})}$$

6 Conclusions

GA-P methods can discover an empirical law from a set of samples. The method is very flexible, because it allows us to select the maximum complexity of the expression, the maximum number of parameters and an arbitrary set of operations. Therefore, GA-P's are very convenient when compared to other methods able to make this kind of study: trial and error, neural networks, classical regression.

In this work we have adapted the GA-P algorithm to fit a model to a set of imprecise examples. The method produces multivalued estimations of the parameters that can be converted into robust punctual estimations and also gives an estimation of a lower bound of the expected number of times the desired output is in the range that the model generates.

Acknowledgments

I would like to thank Oscar Cordón and Francisco Herrera, from Granada University, for applying Wang-Mendel and TSK fuzzy modeling methods to the dataset and for their useful comments about this draft.

References

[1] Bojadziev, G. *Fuzzy Sets, Fuzzy Logic, Applications.* World Scientific. 1995.

[2] Couso, I. *La Envolvente Probabilística. Definición y Propiedades.* Trabajo de Investigación. Universidad de Oviedo. Departamento de Estadística. 1997.

[3] Fonseca, C., Fleming, P.J. "An Overview of Evolutionary Algorithms in Multiobjective Optimization". Evolutionary Computation 3, 1-16. 1995.

[4] Howard, L.; D'Angelo, D. "The GA-P: A Genetic Algorithm and Genetic Programming Hybrid" IEEE Expert. June 1995. 11-15. 1995.

[5] Iba, H.,Sato, T. De Garis, H. "System Identification approach to genetic programming". Proc First IEEE Conf on Evolutionary Computation. 401-406. vol 1. 1994.

[6] *Genetic Programming. On the programming of computers by mean of natural selection.* MIT Press. 1993.

[7] Ishibuchi, H., Tanaka, H., Okada, H. "An architecture of neural networks with interval weights and its application to fuzzy regression analysis". Fuzzy Sets and Systems 57. 27-39. 1993.

[8] Kacprzyk, J. *Fuzzy Regression Analysis*. Omnitech Press, Warsaw. 1992.

[9] Ljung, L. *System Identification: Theory for the User* Prentice Hall. 1987.

[10] Renders, J.M.; Bersini, H. "Hybridizing genetic algorithms with hill-climbing methods for global optimization: two possible ways" Proc. first IEEE Conf. Evolutionary Computation. 312-317, vol. 1. 1994.

[11] Renders, J.M., Flasse, S. P. "Hybrid Methods Using Genetic Algorithms for Global Optimization". IEEE Transactions on SMC. Part B: Cybernetics. Vol 26, NO. 2, April 1996.

[12] Sánchez, L. "A Random Sets Based Method for Identifying Fuzzy Models". To appear in Fuzzy Sets and Systems. 1998.

[13] Shafer, G. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, New Jersey. 1986.