VERSITA

Int. J. Appl. Math. Comput. Sci., 2012, Vol. 22, No. 2, 461–476
DOI: 10.2478/v10006-012-0035-4

amcs

NEURO–ROUGH–FUZZY APPROACH FOR REGRESSION MODELLING FROM MISSING DATA

KRZYSZTOF SIMIŃSKI

Institute of Informatics
Silesian University of Technology, ul. Akademicka 16, 44-100 Gliwice, Poland
e-mail: Krzysztof.Siminski@polsl.pl

Real life data sets often suffer from missing data. The neuro-rough-fuzzy systems proposed hitherto often cannot handle such situations. The paper presents a neuro-fuzzy system for data sets with missing values. The proposed solution is a complete neuro-fuzzy system. The system creates a rough fuzzy model from presented data (both full and with missing values) and is able to elaborate the answer for full and missing data examples. The paper also describes the dedicated clustering algorithm. The paper is accompanied by results of numerical experiments.

Keywords: neuro-fuzzy, ANNBFIS, missing values, marginalisation, imputation, rough fuzzy set, clustering.

1. Introduction

Real life data often lack some values. The reasons for this are diverse: refusal to answer some questions in a questionnaire, inapplicability of questions, irrelevant or unknown attributes, errors in answer acquisition, random noise, impossible values, impossibility to get data (e.g., a patient has died). Missing data are a problem in medical research for two main reasons. The first one is practical impossibility of collecting all the data. The second problem appears when the data are used retrospectively, i.e., the data were gathered for other purposes than the research needs. Renz *et al*. (2002) give an example where only 1 patient out of 55 had all blood tests done. Overall, 9.2% of blood test results are missing. Lakshminarayan *et al*. (1999) present a real life data set with more than 50% of the values missing.

Acuña and Rodriguez (2004) describe the classification of data sets with missing values. The data sets with less than 1% of missing values are labelled as trivial, 1–5% as manageable. For data sets with 5–15% of missing values some sophisticated methods are required, and finally more than 15% missing values "severely impact any kind of interpretation".

Generally, three approaches are used to handle the problem of missing values:

1. Imputation: the unknown values are substituted with estimated ones (Renz *et al*., 2002; Wagstaff and Laidler, 2005; Dempster *et al*., 1977; Ghahramani and Jordan, 1995; Zhang *et al*., 2007; Zhang, 2011).
2. Marginalisation (Whole Data Strategy, WDS): the data tuples with missing values are removed from the data set (Troyanskaya *et al*., 2001; Hathaway and Bezdek, 2001) or the features (attributes) with missing values are ignored (Cooke *et al*., 2001), which leads to the lowering of the task dimensionality.
3. Rough sets express imprecision caused by the lack of data (Nowicki, 2006; Grzymala-Busse and Hu, 2001; Grzymala-Busse, 2006).

The advantage of both data imputation and marginalisation is their simplicity. Imputation is more frequently used than marginalisation (Himmelspach and Conrad, 2010). The results elaborated based on data sets with imputed values cannot be fully trusted (Troyanskaya *et al*., 2001). The imputed values may have no physical meaning in real life (Wagstaff and Laidler, 2005). The missing values are commonly replaced with zeros, random numbers, a mean value over all data set, a mean value over the class the example belongs to, deductive imputation (the missing values are deduced from other information of the pattern), regression-based imputation (Chan *et al*., 1976) or a value based on real distribution (the missing values are replaced with random values with data set distribution) (Wagstaff and Laidler, 2005). The Expectation–Maximisation (EM) (Dempster *et al*., 1977) algorithm is applied by Ghahramani and Jordan (1995). Imputation

based on nearest neighbourhood is proposed by Zhang *et al*. (2007) and Zhang (2011). To avoid imputation of non-existing values, the hot-deck procedure has been proposed (Rubin, 1987) with various distance measures (Fuller and Kim, 2005; Farhangfar *et al*., 2007). The impact of imputation of missing values on the classification error is discussed by Farhangfar *et al*. (2008)

Not many neuro-fuzzy approaches for handling missing data have been proposed (Nowicki, 2006; 2008; 2009; 2010; Korytkowski *et al.*, 2008). These systems are designed for classification. In this paper we propose a system for regression (continuous decision class).

The proposed system is in a certain sense an extension of that presented by Nowicki (2008) with the rough set approach. Between that system and our approach there are two essential differences. Our system is designed for the regression modelling task, not for classification. In the work of Nowicki (2008) the creation of rough fuzzy rules is not described in detail. Ours is a complete system for data sets with missing values. The system can create the fuzzy model based on full or missing value data sets and can elaborate the answer for both full tuples or tuples with missing values. The values in the train data set can be missing from all attributes. The only limitation is that at least one data tuple should be complete with no missing attributes.

Missing data are modelled with a rough fuzzy set approach. In order to create a rough fuzzy model, both marginalisation and imputation techniques are used. The former is used to create the data subset containing data tuples with sure values. The second data set contains the data with imputed values. These two data sets are used for clustering the data. The data are clustered into rough fuzzy (type-2 fuzzy) sets. Based on these clusters, fuzzy rules are extracted and the fuzzy rule base is created. The system produces an answer both for full value data tuples and for tuples with missing values. The answer consists of two values: upper and lower approximation. A general overview of creation of the fuzzy model is presented in Fig. 1.

The paper is organised as follows. Section 2 describes neuro-fuzzy systems, Section 4 presents our proposition (model creation and elaboration of answers). The experiments are described in Section 5. Finally, Section 6 presents the conclusions.

In the paper, the empty characters ($\mathbb{A}$) are used to denote the sets, bolds ($\mathbf{a}$): matrices and vectors, uppercase italics (A): the cardinality of sets, lowercase italics (a): scalars and set elements, bold italics ($\boldsymbol{a}$): relations. A detailed list of symbols is gathered in Table 1.

Input: $\mathbb{X}$—array of tuples
Input: R—number of rules
Output: $\mathcal{M}$—fuzzy model
// preprocessing of data, Sec. 4.1.1:
1 $\underline{\mathbb{X}} \leftarrow$ marginalisation $(\mathbb{X})$;
2 $\overline{\mathbb{X}} \leftarrow$ imputation $(\mathbb{X})$;
// creation of model:
3 $\left[\underset{\sim}{\mathbf{m}}, \widetilde{\mathbf{m}}\right] \leftarrow$ clustering $(\underline{\mathbb{X}}, \overline{\mathbb{X}}, R)$; *// Sec. 4.1.2*
4 $[\mathcal{L}, \mathcal{U}] \leftarrow$ extraction $(\underset{\sim}{\mathbf{m}}, \widetilde{\mathbf{m}}, \underline{\mathbb{X}}, \overline{\mathbb{X}})$; *// Sec. 4.1.3*
5 $[\mathcal{L}, \mathcal{U}] \leftarrow$ tuning $(\mathcal{L}, \mathcal{U}, \underline{\mathbb{X}}, \overline{\mathbb{X}})$; *// Sec. 4.1.4*
6 $\mathcal{M} \leftarrow [\mathcal{L}, \mathcal{U}]$;
7 **return** $\mathcal{M}$;

Fig. 1. Creation of the fuzzy model.

2. Fuzzy inference system with parametrised consequences

The neuro-fuzzy system with parametrised consequences (Czogała and Łęski, 2000; Łęski and Czogała, 1999) is a system combining the Mamdani–Assilan (Mamdani and Assilian, 1975) and the Takagi–Sugeno–Kang (Takagi and Sugeno, 1985; Sugeno and Kang, 1988) approach. The fuzzy sets in consequences are isosceles triangles (as in the Mamdami–Assilan system), but are not fixed—their location is calculated as a linear combination of attribute values as the localisation of singletons in Takagi–Sugeno–Kang system.

The idea of the system with parametrised consequences is presented in Fig. 2. The figure is taken after Czogała and Łęski (2000) with modifications.

The system with parametrised consequences is a MISO system. The rule base $\mathbb{R}$ contains fuzzy rules r in form of fuzzy implications

$$r : \mathbf{x} \text{ is } \mathfrak{a} \rightsquigarrow y \text{ is } \mathfrak{b}, \tag{1}$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_A]^{\mathrm{T}}$ and y are linguistic variables, $\mathfrak{a}$ and $\mathfrak{b}$ are fuzzy linguistic terms (values).

The linguistic variable $\mathfrak{a}$ (in the rule premise) is represented in the system as a fuzzy set $\mathbb{A}$ in an A-dimensional space. In each dimension the set $\mathbb{A}$ is described with the Gaussian membership function:

$$\mu_a(x_a) = \exp\Big(-\frac{(x_a - c_a)^2}{2s_a^2}\Big), \tag{2}$$

where c_a is the core location for the a-th attribute and s_a is this attribute Gaussian bell deviation. The membership of the variable $\mathbf{x}$ to the fuzzy set $\mathbb{A}^{(r)}$ in the r-th rule is defined as a T-norm:

$$\begin{aligned}\mu_{\mathbb{A}^{(r)}}(\mathbf{x}) &= \mu_{a_1^{(r)}}(x_1) \star \cdots \star \mu_{a_A^{(r)}}(x_A) \\ &= \operatorname*{\bigstar}_{a \in \mathbb{A}} \mu_{a^{(r)}}(x_a),\end{aligned} \tag{3}$$

Table 1. Symbols and abbreviations

$\mathbb{R}$	set of rules, rule base
r	rule, $r \in \mathbb{R}$
R	number of rules, $R = \|\mathbb{R}\|$; number of clusters
$\mathfrak{a}, \mathfrak{b}$	fuzzy linguistic terms
$\mathbb{X}$	set of tuples, data examples
$\mathbf{x}$	tuple, data example, $\mathbf{x} \in \mathbb{X}$
$\mathbf{x}_i$	i-th tuple
x	descriptor of tuple, $\mathbf{x} = [x_1, \ldots, x_A]^{\mathrm{T}}$
X	number of tuples, $X = \|\mathbb{X}\|$
X_u	number of tuples in upper set, $X_u = \|\overline{\mathbb{X}}\|$
X_l	number of tuples in lower set, $X_l = \|\underline{\mathbb{X}}\|$
$\mathbb{A}$	set of attributes
a	attribute, $a \in \mathbb{A}$
A	number of attributes in tuple, $\mathbb{A} = \|\mathbb{A}\|$
A_t	threshold number of attributes
$\mathbf{m}$	partition matrix
$\widetilde{m}_{ru}$	membership value of u-th tuple to r-th "upper" cluster
$\underset{\sim}{m}_{rl}$	membership value of l-th tuple to r-th "lower" cluster
$\mu_{\mathbb{A}}(a)$	membership value of element a to set $\mathbb{A}$
d_{rj}	distance between r-th cluster's centre and j-th tuple
η_u	weight of u-th tuple in upper set
$\mathbf{v}_r$	centre of r-th cluster
f	fuzzification parameter, here $f = 2$
$\mathcal{U}$	model based on $\overline{\mathbb{X}}$ data set
$\mathcal{L}$	model based on $\underline{\mathbb{X}}$ data set
$\mathbb{B}$	triangle set in consequence
$\mathbb{B}'$	fuzzy set of rule implication
$\boldsymbol{q}$	relation
$\rightsquigarrow$	fuzzy implication
$\star$	T-norm

where $\star$ denotes the T-norm and $\mathbb{A}$ stands for the set of attributes. The membership to the r-th fuzzy set $\mathbb{A}^{(r)}$ (the premise of the r-th fuzzy rule) is simultaneously the firing strength $F^{(r)}$ of the r-th rule,

$$F^{(r)}(\mathbf{x}) = \mu_{\mathbb{A}^{(r)}}(\mathbf{x}). \tag{4}$$

The term $\mathfrak{b}$ (in the rule consequence) is represented by an isosceles triangle fuzzy set $\mathbb{B}$ with the base width w, the altitude of the triangle equal to 1. The localisation of the core of the triangle membership function is determined by linear combination of input attribute values:

$$\begin{aligned} y^{(r)}(\mathbf{x}) &= \mathbf{p}^{\mathrm{T}} \cdot \left[1, \mathbf{x}^{\mathrm{T}}\right]^{\mathrm{T}} \\ &= \left[p_0^{(r)}, p_1^{(r)}, \ldots, p_A^{(r)}\right] \cdot [1, x_1, \ldots, x_A]^{\mathrm{T}}, \end{aligned} \tag{5}$$

where $\mathbf{p}$ is the parameter vector of the consequence.

The output of the rule is the fuzzy value of the fuzzy implication,

$$\mu_{\mathbb{B}'^{(r)}}(\mathbf{x}) = \mu_{\mathbb{A}^{(r)}}(\mathbf{x}) \rightsquigarrow \mu_{\mathbb{B}^{(r)}}(\mathbf{x}), \tag{6}$$

where the squiggle arrow ($\rightsquigarrow$) stands for the fuzzy implication. The shape of the fuzzy set $\mathbb{B}'^{(r)}$ depends on the fuzzy implication used (Czogała and Łęski, 2000). The answers $\mu_{\mathbb{B}'^{(r)}}$ of all R rules are then aggregated into one fuzzy answer of the system,

$$\mu_{\mathbb{B}'}(\mathbf{x}) = \bigoplus_{r=1}^{R} \mu_{\mathbb{B}'^{(r)}}(\mathbf{x}). \tag{7}$$

In order to get a crisp answer y_0, the fuzzy set $\mathbb{B}'$ is defuzzified with the MICOG method (Czogała and Łęski, 2000). This approach removes the non-informative parts of the aggregated fuzzy sets and takes into account only the informative ones. The aggregation and defuzzyfication may be quite expensive, but it has been proved (Czogała and Łęski, 2000) that the crisp system output can be expressed as

$$y_0 = \frac{\sum_{r=1}^{R} g^{(r)}(\mathbf{x})\, y^{(r)}(\mathbf{x})}{\sum_{r=1}^{R} g^{(r)}(\mathbf{x})}. \tag{8}$$

The function g depends on the fuzzy implication; in the system the Reichenbach one is used, so for the r-th rule the function g is

$$g^{(r)}(\mathbf{x}) = \frac{w^{(r)}}{2} F^{(r)}(\mathbf{x}). \tag{9}$$

The forms of the g function for various implications can be found in the original work introducing the ANNBFIS system (Czogała and Łęski, 2000). Some inaccuracies are discussed by Nowicki (2006) and Łęski (2008).

Creation of the fuzzy model $\mathcal{M}$ is done in three steps: clustering of the input domain, extraction of rules' premises and tuning of the rules (this step is also responsible for creation of rules consequences) (Czogała and Łęski, 2000).

3. Rough sets

Rough sets were proposed by Pawlak (1982). The equivalence relation $\boldsymbol{q}$ splits the universe set $\mathbb{U}$ into disjoint subsets—equivalence classes. The set $\mathbb{A}$ can be approximated with equivalence classes by means of lower $\underline{\boldsymbol{q}}\mathbb{A}$ and upper $\overline{\boldsymbol{q}}\mathbb{A}$ approximations defined respectively as

$$\underline{\boldsymbol{q}}\mathbb{A} = \{x \in \mathbb{U} : [x]_{\boldsymbol{q}} \subseteq \mathbb{A}\} \tag{10}$$

and

$$\overline{\boldsymbol{q}}\mathbb{A} = \{x \in \mathbb{U} : [x]_{\boldsymbol{q}} \cap \mathbb{A} \neq \varnothing\}, \tag{11}$$

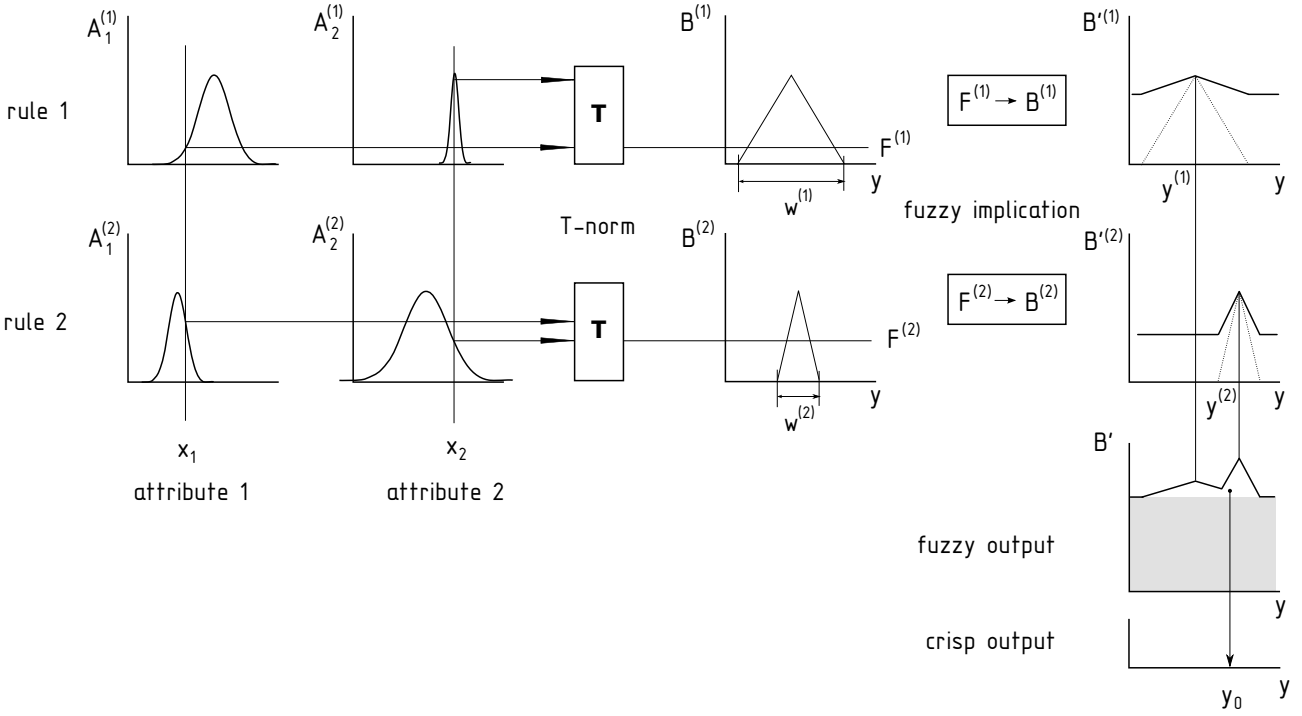

Fig. 2. Schema of the neuro-fuzzy system with parametrised consequences. The input has two attributes and the rule base is composed of two fuzzy rules. The premises of the rules are responsible for determining the firing strength of the rules. The firing strength is the left operand of the fuzzy implication. The right hand operand is the $\mathbb{B}$ fuzzy triangle set, the location of which is determined with the formula (5). The result of the r-th fuzzy implication is the fuzzy set $\mathbb{B}'^{(r)}$. The fuzzy results of the implications are then aggregated. The non-informative part (grey rectangular in the picture) is thrown away in aggregation. The informative part (white mountain-like part of the $\mathbb{B}'$ set) is then defuzzyfied with the centre of gravity method. The answer of the system is the crisp number y_0.

where $[x]_{\boldsymbol{q}}$ is an equivalence class defined as

$$[x]_{\boldsymbol{q}} = \{a \in \mathbb{U} : x \ \boldsymbol{q} \ a\}. \tag{12}$$

The lower and upper approximations build up the rough set $\langle \underline{\boldsymbol{q}}\mathbb{A}, \overline{\boldsymbol{q}}\mathbb{A}\rangle$. The lower approximation $\underline{\boldsymbol{q}}\mathbb{A}$ of the set $\mathbb{A}$ is a set of subsets that without doubt belong to the set $\mathbb{A}$. The upper approximation $\overline{\boldsymbol{q}}\mathbb{A}$ is a set of subsets that have some nonempty element common with the set $\mathbb{A}$. It is worth mentioning that

$$\underline{\boldsymbol{q}}\mathbb{A} \subseteq \mathbb{A} \subseteq \overline{\boldsymbol{q}}\mathbb{A}. \tag{13}$$

The rough set is a good instrument for handling uncertainty. The following statements are true:

$$a \in \underline{\boldsymbol{q}}\mathbb{A} \Rightarrow a \in \mathbb{A} \tag{14}$$

and

$$a \notin \overline{\boldsymbol{q}}\mathbb{A} \Rightarrow a \notin \mathbb{A}. \tag{15}$$

When $a \in \overline{\boldsymbol{q}}\mathbb{A} \wedge a \notin \underline{\boldsymbol{q}}\mathbb{A}$, we cannot say for sure that the element a belongs or does not belong to set $\mathbb{A}$.

The concept of joining rough and fuzzy sets comes from Dubois and Prade (1990). Two ways of combining two kinds of sets were proposed: rough fuzzy sets (lower and upper approximations of fuzzy sets are defined) and fuzzy rough sets (lower and upper approximations of fuzzy sets are fuzzy). In our approach we use rough fuzzy sets.

For simpler notation the relation $\boldsymbol{q}$ will be omitted and instead of $\overline{\boldsymbol{q}}\mathbb{A}$ and $\underline{\boldsymbol{q}}\mathbb{A}$ we will use $\overline{\mathbb{A}}$ and $\underline{\mathbb{A}}$, respectively.

4. Our approach

The drawback of the methods for handling missing values mentioned in Introduction is no distinction between original untouched data and imputed values (Himmelspach and Conrad, 2010). Further, the imputed values may have no medical/physical meaning (Wagstaff and Laidler, 2005), thus the models based on imputed data cannot be fully trusted (Troyanskaya *et al*., 2001). The method proposed by Wagstaff (2004) as well as Wagstaff and Laidler (2005) divides the feature set $\mathbb{F}$ into features $\mathbb{F}_o$ with no lacking values and partially observed features $\mathbb{F}_m$. Thus the algorithm cannot be used when the values are missing from all attributes ($\mathbb{F}_o = \varnothing$).

In our solution both approaches are used: marginalisation and imputation. The former removes the tuples

with missing values. The remaining tuples contain only original data. This data set $\underline{\mathbb{X}}$ is used to create the lower rough set cluster—the core cluster. The latter is used to handle data with missing values. All data augmented with imputed data stand for the upper data set $\overline{\mathbb{X}}$ and are used to elaborate the upper rough set cluster—the "cloud" cluster containing the core cluster. The lower data set is a subset of the upper data set: $\underline{\mathbb{X}} \subseteq \overline{\mathbb{X}}$. This approach maintains the distinction between original and imputed values. If the data set lacks no values, the upper and lower data sets are the same: $\underline{\mathbb{X}} = \overline{\mathbb{X}} = \mathbb{X}$.

4.1. Model creation. The creation of the model comprises four steps: preprocessing of data, clustering, extraction of rules and tuning. The following subsections describe these.

4.1.1. Preprocessing of data. Preprocessing of data leads to creation of data sets $\underline{\mathbb{X}}$ and $\overline{\mathbb{X}}$. The former is created with marginalisation, the latter with imputation.

Marginalisation. The tuples with missing values are excluded from the data set $\underline{\mathbb{X}}$. This set contains only tuples $\mathbf{x} \in \mathbb{X}$ that lack no values. Marginalisation excludes the tuples, not the attributes, thus there is no dimensionality reduction. This approach is similar to one used by Troyanskaya *et al*. (2001), as well as Hathaway and Bezdek (2001).

Imputation into the upper data set. The tuples with missing values are substituted with new tuples with imputed values. If the tuple lacks n values, it is substituted with k^n tuples with all combinations of imputed values (these are the mean values m of the missing attribute calculated from values existing in other tuples, $m+\sigma$, where σ is the standard deviation of the attribute, $m-\sigma$, thus $k=3$). The maximum and minimum values are not used here, because the extreme values may be outliers and one extreme bias value can substantially influence the clustering process.

Unfortunately, the number of new tuples grows very fast with the number n of missing values from the original tuple. This explosion in the number of tuples can have disadvantageous influence on the efficacy of calculations. Thus when the tuple lacks more a A_t values, not all possible combinations are imputed, but for each missing value v, k tuples are imputed and other missing attributes $q \neq v$ are imputed with means of the respective attributes. So only kn new tuples are added.

Figure 3 presents an example of a data set with missing values denoted with question marks. If $A_t \geq 2$, the tuple $\mathbf{x}_1$ will be substituted with $k^n = 3^2 = 9$ tuples (Fig. 4). If $A_t \geq 2$, the tuple in question will be imputed with $kn = 3 \cdot 2 = 6$ tuples (Fig. 5). The twofold approach is used because in a real-life data set the tuple may lack 8 or more values.

	a_1	a_2	a_3	a_4
$\mathbf{x}_1$	2	?	?	1
$\mathbf{x}_2$	5	2	8	4
$\mathbf{x}_3$	1	2	9	2
$\mathbf{x}_4$	4	5	7	2
$\mathbf{x}_5$	2	5	6	1
$\mathbf{x}_6$	3	0	5	3
average		2.80	7.00	
st. dev.		2.17	1.58	

Fig. 3. Example of a data set with missing values (denoted with exclamation marks). The last two rows show the average values and standard deviation of attributes a_2 and a_3.

	a_1	a_2	a_3	a_4
$\mathbf{x}_1$	2	?	?	1
$\mathbf{x}_7$	2	0.63	5.42	1
$\mathbf{x}_8$	2	0.63	7.00	1
$\mathbf{x}_9$	2	0.63	8.58	1
$\mathbf{x}_{10}$	2	2.80	5.42	1
$\mathbf{x}_{11}$	2	2.80	7.00	1
$\mathbf{x}_{12}$	2	2.80	8.58	1
$\mathbf{x}_{13}$	2	4.97	5.42	1
$\mathbf{x}_{14}$	2	4.97	7.00	1
$\mathbf{x}_{15}$	2	4.97	8.58	1

Fig. 4. Tuples nos. 7–15 are imputed in the data set from Fig. 3 in place of tuple $\mathbf{x}_1$ when $A_t \geq n = 2$.

If the tuple with missing values is substituted with t imputed tuples, each of these imputed tuples is assigned the weight $\eta = 1/t$. The weight is treated as a condition in conditional FCM clustering.

4.1.2. Clustering. Both data sets $\underline{\mathbb{X}}$ and $\overline{\mathbb{X}}$ are used in clustering. The clustering divides the input domain into regions that are then converted into rule premises.

For clustering, the hybrid FCM algorithm is used. For the data set $\underline{\mathbb{X}}$, the standard FCM algorithm (Dunn, 1973) is employed. For the data set $\overline{\mathbb{X}}$, the conditional FCM proposed by Pedrycz (1998) is applied. One more modification is used. The "upper" and "lower" clusters are gathered into pairs with common cluster centres. The "lower" data set is created upon only original complete data values (set $\underline{\mathbb{X}}$). This data set is more reliable than the "upper" data set with imputed values. Thus the cluster centres are elaborated based only on lower data set membership functions. The clustering is based on minimising the criterion function J (for the description of symbols,

	a_1	a_2	a_3	a_4
$\mathbf{x}_1$	2	?	?	1
$\mathbf{x}_{16}$	2	2.80	5.42	1
$\mathbf{x}_{17}$	2	2.80	7.00	1
$\mathbf{x}_{18}$	2	2.80	8.58	1
$\mathbf{x}_{19}$	2	0.63	7.00	1
$\mathbf{x}_{20}$	2	2.80	7.00	1
$\mathbf{x}_{21}$	2	4.97	7.00	1

Fig. 5. Tuples no. 16–21 are imputed in the data set from Fig. 3 in place of tuple $\mathbf{x}_1$ when $A_t < n = 2$.

see Table 1):

$$J = \sum_{r=1}^{R} \left[\sum_{u=1}^{X_u} \widetilde{m}_{ru}^{f} d_{ru}^2 + \sum_{l=1}^{X_l} \underset{\sim}{m}_{rl}^{f} d_{rl}^2 \right] \tag{16}$$

with the conditional boundary

$$\underset{u \in \overline{\mathbb{X}}}{\forall} \sum_{r=1}^{R} \widetilde{m}_{ru} = \eta_u, \tag{17}$$

where η_u is a tuple's weight (cf. Section 4.1.1). Owing to this boundary, the tuples with imputed values (as being less reliable) have lesser influence on the results of clustering. For lower clustering, the standard FCM boundary is applied because only full tuples (with no imputed values) are clustered:

$$\underset{l \in \underline{\mathbb{X}}}{\forall} \sum_{r=1}^{R} \underset{\sim}{m}_{rl} = 1. \tag{18}$$

The cluster centres are elaborated based only on "lower" (more reliable) membership values:

$$\mathbf{v}_r = \frac{\sum\limits_{l=1}^{X_l} \underset{\sim}{m}_{rl} \mathbf{x}_l}{\sum\limits_{l=1}^{X_l} \underset{\sim}{m}_{rl}}. \tag{19}$$

For elaborating formulae for modification of membership values, it is common to use Lagrange multipliers. The function L is defined as

$$\begin{aligned} &L(\widetilde{\mathbf{m}}, \underset{\sim}{\mathbf{m}}, \lambda_1, \lambda_2) \\ &= \sum_{u=1}^{X_u} \sum_{r-1}^{R} \widetilde{m}_{ru}^{f} d_{ru}^2 + \sum_{l=1}^{X_l} \sum_{r-1}^{R} \underset{\sim}{m}_{rl}^{f} d_{rl}^2 \\ &\quad - \lambda_1 \left(\sum_{r-1}^{R} \widetilde{m}_{ru} - \eta_u \right) - \lambda_2 \left(\sum_{r-1}^{R} \underset{\sim}{m}_{rl} - 1 \right), \end{aligned} \tag{20}$$

where $\widetilde{\mathbf{m}}$ and $\underset{\sim}{\mathbf{m}}$ are "upper" and "lower" partition matrices, respectively. It should be clearly stated that these symbols do not represent upper and lower rough approximations. These are the partition matrices used to calculate the rough approximation of cluster fuzziness parameters $\underline{\mathbf{s}}$ and $\overline{\mathbf{s}}$. To express this difference, a tilde is used as a diacritic instead of a bar.

The symbols λ_1 and λ_2 stand for Lagrange multipliers and d_{rq} is a distance between the q-th datum from the q-th cluster centre,

$$d_{rj}^2 = (\mathbf{v}_r - \mathbf{x}_q)^{\mathrm{T}} \mathbf{A} (\mathbf{v}_r - \mathbf{x}_q), \tag{21}$$

where $\mathbf{A}$ is positive-defined matrix. The derivatives of L (Eqn. 20) are

$$\underset{1 \le u \le X_u}{\forall} \underset{1 \le s \le R}{\forall} \frac{\partial L}{\partial \widetilde{m}_{su}} = f \widetilde{m}_{su}^{f-1} d_{su}^2 - \lambda_1 = 0, \tag{22}$$

$$\begin{aligned} f \widetilde{m}_{su}^{f-1} d_{su}^2 - \lambda_1 &= 0, \\ f \widetilde{m}_{su}^{f-1} d_{su}^2 &= \lambda_1, \\ \widetilde{m}_{su}^{f-1} d_{su}^2 &= \frac{\lambda_1}{f}, \\ \widetilde{m}_{su}^{f-1} &= \frac{\lambda_1}{f} d_{su}^{-2}, \\ \widetilde{m}_{su} &= \left(\frac{\lambda_1}{f} \right)^{\frac{1}{f-1}} d_{su}^{\frac{2}{1-f}}. \end{aligned} \tag{23}$$

Substituting Eqn. (23) into Eqn. (17) we get

$$\begin{aligned} \sum_{r-1}^{R} \left(\frac{\lambda_1}{f} \right)^{\frac{1}{f-1}} d_{ru}^{\frac{2}{1-f}} &= \eta_u, \\ \left(\frac{\lambda_1}{f} \right)^{\frac{1}{f-1}} \sum R_{r-1} d_{ru}^{\frac{2}{1-f}} &= \eta_u. \end{aligned} \tag{24}$$

Combining Eqns. (23) and (24) we obtain

$$\widetilde{m}_{su} = \frac{\eta_u d_{su}^{\frac{2}{1-f}}}{\sum\limits_{r-1}^{R} d_{ru}^{\frac{2}{1-f}}}. \tag{25}$$

Analogously, from

$$\underset{1 \le l \le X_l}{\forall} \underset{1 \le s \le I}{\forall} \frac{\partial L}{\partial \widetilde{m}_{sl}} = f \underset{\sim}{m}_{sl}^{f-1} d_{sl}^2 - \lambda_2 = 0 \tag{26}$$

we get

$$\underset{\sim}{m}_{sl} = \left(\frac{\lambda_2}{f} \right)^{\frac{1}{f-1}} d_{sl}^{\frac{2}{1-f}}. \tag{27}$$

Substituting Eqn. (27) in Eqn. (18),

$$\left(\frac{\lambda_2}{f} \right)^{\frac{1}{f-1}} \sum_{r-1}^{R} d_{rl}^{\frac{2}{1-f}} = 1, \tag{28}$$

and combining Eqns. (27) and (28), we obtain

$$\underset{\sim}{m}_{sl} = \frac{d_{sl}^{\frac{2}{1-f}}}{\sum\limits_{r-1}^{R} d_{rl}^{\frac{2}{1-f}}}. \tag{29}$$

This clustering algorithm creates clusters that type-2 fuzzy sets. Type-2 fuzzy clustering is not widely used. Hwang and Rhee (2004) propose a clustering algorithm that is a modification of the FCM algorithm. The two membership functions are achieved by applying various values of the f parameter. The values of the f parameters are selected manually by the user and are not tuned nor modified during the clustering procedure. In our approach, the f parameters for both fuzzy sets are constant ($f = 2$) and the fuzzy sets differ by the parameter s. The gap between "upper" and "lower" fuzzy sets is fitted automatically and does not have to be set manually. The cluster represented by a pair of fuzzy sets can also be regarded as a rough fuzzy set. The examples of clusters are presented in Fig. 6

4.1.3. Extraction of clusters from partition matrices. When the clustering is finished, the clusters are transformed into rule premises with the method described by Czogała and Łęski (2000), as well as Łęski (2008). Based on partition matrices $\underset{\sim}{\mathbf{m}}$ and $\widetilde{\mathbf{m}}$, the parameters c, s (cf. Eqn. (2)) are calculated. The cluster centre $\mathbf{c} = [c_1, c_2, \ldots, c_A]$ is calculated with Eqn. (19) for both upper and lower fuzzy sets, thus $\mathbf{c} = \underline{\mathbf{c}} = \overline{\mathbf{c}} = \mathbf{v}$. The cluster centres become cores of the Gaussian function (Eqn. (2)). The parameter $\mathbf{s}$ for the r-th rule is elaborated with the formula

$$\overline{\mathbf{s}}_r = \sqrt{\frac{\sum\limits_{u=1}^{X_u} \widetilde{m}_{ru}^f \left(\mathbf{x}_u - \mathbf{v}_r\right)^2}{\sum\limits_{u=1}^{X_u} \widetilde{m}_{ru}^f}} \tag{30}$$

for upper clusters and

$$\underline{\mathbf{s}}_r = \sqrt{\frac{\sum\limits_{l=1}^{X_l} \underset{\sim}{m}_{rl}^f \left(\mathbf{x}_l - \mathbf{v}_r\right)^2}{\sum\limits_{l=1}^{X_l} \underset{\sim}{m}_{rl}^f}} \tag{31}$$

for lower ones. The above formulae approximate the standard deviation of the Gauss function expressed by Eqn. (2). An example of interpretation of $\overline{s}$ and $\underline{s}$ is presented in Fig. 6.

The elaborated values of $\underline{\mathbf{s}} = [\underline{s}_1, \ldots, \underline{s}_A]$, $\overline{\mathbf{s}} = [\overline{s}_1, \ldots, \overline{s}_A]$ and $\mathbf{c} = [c_1, \ldots, c_A]$ enable the calculation of memberships $\underline{\mu}, \overline{\mu}$ to the rough fuzzy set representing the attribute in the rule premise with the formula (2). For each attribute a and for each data tuple $\mathbf{x}$, the following relation is true:

$$\underset{a \in \mathbb{A}}{\forall} \quad \underset{\mathbf{x} \in \mathbb{X}}{\forall} \quad \underline{\mu}_{a^{(r)}}(\mathbf{x}) \le \overline{\mu}_{a^{(r)}}(\mathbf{x}). \tag{32}$$

4.1.4. Tuning. The above stages lead to elaboration of rule premises. The next procedure is tuning. It has two aims. The first one is better fitting of the model to the presented data. The values of $\mathbf{c}$, $\underline{\mathbf{s}}$ and $\overline{\mathbf{s}}$ extracted from the clustering of the input domain are tuned to fit the presented data. The second aim of tuning is the elaboration of linear parameters $\mathbf{p}$ in the conclusion (cf. Eqn. (5)).

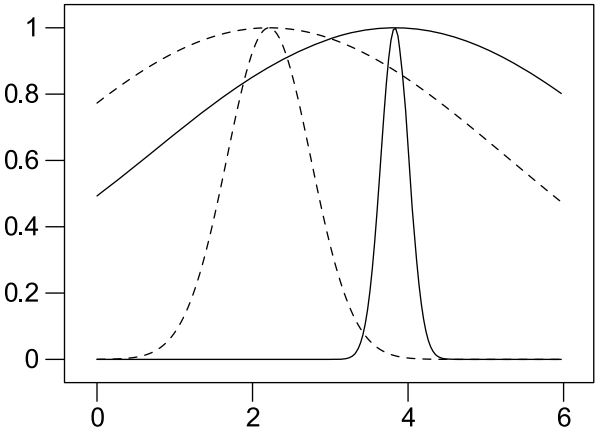

Fig. 6. Example of membership functions of one-dimensional clusters extracted with rough clustering. The parameters of the cluster pairs are denoted as $c_{\underline{s}}^{\overline{s}} = 3.83_{0.18}^{3.22}$ (solid) and $c_{\underline{s}}^{\overline{s}} = 2.21_{0.54}^{3.08}$ (dashed).

The models $\mathcal{U}$ (created from $\overline{\mathbb{X}}$) and $\mathcal{L}$ (created from $\underline{\mathbb{X}}$) are then tuned with two methods:

1. The premises of the rules and the width $w^{(r)}$ of bases of triangle fuzzy sets $\mathbb{B}^{(r)}$ in consequences are tuned with the gradient method.

2. The parameters $\mathbf{p}$ for calculation of localisations of consequences are elaborated with the linear regression, least mean square method.

Similarly as during clustering, the centres of clusters are tuned based on the "lower" model $\mathcal{L}$. Other parameters of $\mathcal{L}$ are tuned with $\underline{\mathbb{X}}$ and the parameters of $\mathcal{U}$ are tuned respectively with $\overline{\mathbb{X}}$ data sets.

The tuning procedure finishes the elaboration of the models.

4.2. Elaboration of the answer. The models can be created with a full data set or a data set with missing values. The tuple for which we want the system to answer may also be full and complete or with missing values.

If the train data set contains incomplete data tuples, the two models are created and for each complete data tuple two answers are elaborated. The rules in $\mathcal{U}$ have more fuzzy premises, so the membership value for the same tuples is higher in the $\mathcal{U}$ model than in $\mathcal{L}$, because of the feature expressed by the formula (32). The model $\mathcal{U}$ is responsible for calculating $\overline{\mu}_{a^{(r)}}(\mathbf{x})$ and $\mathcal{L}$ for $\underline{\mu}_{a^{(r)}}(\mathbf{x})$. But the values of the answer elaborated by the model are based both on rule premises and consequences, and we

cannot say that the answer of $\mathcal{U}$ is $\overline{y}$ and the answer of $\mathcal{L}$ is $\underline{y}$. Our aim is to calculate the upper and lower boundaries so we take the larger value for $\overline{y}$ and the smaller one for $\underline{y}$:

$$\overline{y}(\mathbf{x}) = \max\left(\mathcal{L}(\mathbf{x}), \mathcal{U}(\mathbf{x})\right), \tag{33}$$

$$\underline{y}(\mathbf{x}) = \min\left(\mathcal{L}(\mathbf{x}), \mathcal{U}(\mathbf{x})\right). \tag{34}$$

When the data set lacks no values, then

$$\overline{y}(\mathbf{x}) = \underline{y}(\mathbf{x}). \tag{35}$$

The elaboration of the system answer for the presented data tuple with missing values requires calculation of values of both rule premise and consequence.

4.2.1. Calculation of the rule premise. The value of the rule premise is also its firing strength. To elaborate the value of the firing strength $F^{(r)}(\mathbf{x}) = \mu_{\mathbb{A}^{(r)}}(\mathbf{x})$ of the r-th rule, the membership to each a-th attribute $\mathbb{A}_a^{(r)}(x_a)$ has to be estimated (cf. Eqn. (3)). If the value of the attribute is missing, it is impossible to apply the formulae (2) and (3) directly. Thus the following procedure is used. The set of attributes is split into the set of present attributes $\mathbb{A}_p$ and absent ones $\mathbb{A}_a$. This division may be different for each tuple. The missing value of membership is substituted with minimal and maximal values of membership:

$$\overline{\mu}_a^{(r)} = \max_{\mathbf{x}\in\mathbb{X}} \overline{\mu}_{a^{(r)}}\left(x_a\right), \tag{36}$$

$$\underline{\mu}_a^{(r)} = \min_{\mathbf{x}\in\mathbb{X}} \underline{\mu}_{a^{(r)}}\left(x_a\right). \tag{37}$$

Thus the formula for calculating the firing strength (Eqn. (3)) is replaced with

$$\overline{\mu}_{\mathbb{A}^{(r)}}(\mathbf{x}) = \left(\underset{a_p\in\mathbb{A}_p}{\bigstar} \overline{\mu}_{a_p^{(r)}}\left(x_{a_p}\right)\right) \star \left(\underset{a_a\in\mathbb{A}_a}{\bigstar} \overline{\mu}_{a_a^{(r)}}\right) \tag{38}$$

and

$$\underline{\mu}_{\mathbb{A}^{(r)}}(\mathbf{x}) = \left(\underset{a_p\in\mathbb{A}_p}{\bigstar} \underline{\mu}_{a_p^{(r)}}\left(x_{a_p}\right)\right) \star \left(\underset{a_a\in\mathbb{A}_a}{\bigstar} \underline{\mu}_{a_a^{(r)}}\right). \tag{39}$$

If the tuple lacks no values, the set $\mathbb{A}_a$ is empty and Eqns. (38) and (39) reduce into Eqn. (3). For the data tuple with missing values, Eqn. (4) splits into two formulae:

$$\overline{F}^{(r)}(\mathbf{x}) = \overline{\mu}_{\mathbb{A}^{(r)}}(\mathbf{x}) \tag{40}$$

and

$$\underline{F}^{(r)}(\mathbf{x}) = \underline{\mu}_{\mathbb{A}^{(r)}}(\mathbf{x}). \tag{41}$$

The next step is to calculate the upper $\overline{y}_0(\mathbf{x})$ and lower $\underline{y}_0(\mathbf{x})$ answers of the system for the $\mathbf{x}$ tuple. The crisp output of the system is calculated with Eqn. (8). Let us rewrite this equation to explicitly denote that the function g is the function of the firing strength (cf. Eqn. (9)):

$$y_0 = \frac{\sum_{r=1}^{R} g^{(r)}\left(F^{(r)}(\mathbf{x})\right) y^{(r)}(\mathbf{x})}{\sum_{r=1}^{R} g^{(r)}\left(F^{(r)}(\mathbf{x})\right)}. \tag{42}$$

From the equation above it is obvious that for calculating $\overline{y}_0$ the upper values $\overline{y}^{(r)}$ should be used. But the question is which value, $\overline{F}^{(r)}(\mathbf{x})$ or $\underline{F}^{(r)}(\mathbf{x})$, should be chosen.

Calculation of maximum and minimum of the function expressed with the formula (42) with respect to firing strength values $F^{(r)}$ is difficult. Thus the approach proposed by Nowicki (2008) will be further developed here.

For choosing which value, $\overline{F}^{(r)}(\mathbf{x})$ or $\underline{F}^{(r)}(\mathbf{x})$, should be used we assume that for a certain value of F the function achieves the suboptimal value. In search for $\overline{y}_0$, if the value of the derivative $\partial y_0/\partial F^{(r)}(\mathbf{x})$ with respect to $F^{(r)}$ is positive, the higher value of the parameter should be used. Thus the upper value $\overline{F}^{(r)}$ is used. Otherwise, the lower value $\underline{F}^{(r)}$ is chosen. A similar situation is when $\underline{y}_0$ is to be calculated. If the derivative of y_0 with respect to $F^{(r)}$ is positive, the lower value $\underline{F}^{(r)}$ is used; otherwise, the upper one $\overline{F}^{(r)}$ is employed.

The calculation of the derivative $\partial y_0/\partial F^{(r)}(\mathbf{x})$, for the dimensions of formulae, is presented in Eqn. (43).

One of the features of the g function is

$$\frac{\partial g\left(F^{(r)}(\mathbf{x})\right)}{\partial F^{(r)}(\mathbf{x})} \geq 0, \tag{44}$$

and the value of $g\left(F^{(q)}(\mathbf{x})\right) \geq 0$ for the Reichenbach implication[1] used in our system thus the sign of the derivative (43) depends on the sign of the sum

$$\sum_q \left(y^{(r)} - y^{(q)}\right). \tag{45}$$

Let Φ denote the above sum. Then

$$\operatorname{sgn}\frac{\partial y_0}{\partial F^{(r)}(\mathbf{x})} = \operatorname{sgn}\sum_q \left(y^{(r)} - y^{(q)}\right) = \operatorname{sgn}\Phi^{(r)}. \tag{46}$$

To calculate $\overline{y}_0$ for each rule r, the factor $\Phi^{(r)}$ is calculated and, if $\Phi^{(r)} \geq 0$, then for this rule in Eqn. (42) the value $\overline{F}^{(r)}$ is used. Otherwise, $\underline{F}^{(r)}$ is applied. To calculate $\underline{y}_0$, if $\Phi^{(r)} \geq 0$, then $\underline{F}^{(r)}$ is used, otherwise $\overline{F}^{(r)}$.

A remark should be now expressed. In calculation of the Φ value for $\partial\overline{y}_0/\partial F^{(r)}(\mathbf{x})$ the difference $\overline{y}^{(r)} - \overline{y}^{(q)}$

[1] This is also true for the Łukasiewicz, Fodor, Kleene–Dienes, Zadeh, Goguen, Gödel and Rescher implications (Łęski, 2008).

$$\frac{\partial y_0}{\partial F^{(r)}(\mathbf{x})} = \frac{1}{\left[\sum_q g\left(F^{(q)}(\mathbf{x})\right)\right]^2} \cdot \left[y^{(q)} \frac{\partial g\left(F^{(r)}(\mathbf{x})\right)}{\partial F^{(r)}(\mathbf{x})} \sum_q g\left(F^{(q)}(\mathbf{x})\right) - \frac{\partial g\left(F^{(r)}(\mathbf{x})\right)}{\partial F^{(r)}(\mathbf{x})} \sum_q y^{(q)} g\left(F^{(q)}(\mathbf{x})\right) \right]$$

$$= \frac{\frac{\partial g\left(F^{(r)}(\mathbf{x})\right)}{\partial F^{(r)}(\mathbf{x})}}{\left[\sum_q g\left(F^{(q)}(\mathbf{x})\right)\right]^2} \cdot \left[y^{(r)} \sum_q g\left(F^{(q)}(\mathbf{x})\right) - \sum_q y^{(q)} g\left(F^{(j)}(\mathbf{x})\right) \right]$$

$$= \frac{\frac{\partial g\left(F^{(r)}(\mathbf{x})\right)}{\partial F^{(r)}(\mathbf{x})}}{\left[\sum_q g\left(F^{(q)}(\mathbf{x})\right)\right]^2} \cdot \sum_q \left(y^{(r)} - y^{(q)}\right) g\left(F^{(q)}(\mathbf{x})\right) \tag{43}$$

Table 2. Comparison of two approaches for classification. The symbols $\overline{\tilde{D}}A^r$ and $\underline{\tilde{D}}A^r$ are defined in the same way as $\overline{F}$ and $\underline{F}$, respectively.

	Nowicki	our approach
$\overline{y}_0$	$\overline{z}_j^r = 1$ use $\overline{\tilde{D}}A^r$ $\overline{z}_j^r = 0$ use $\underline{\tilde{D}}A^r$	$y^{(r)} = 1 \Rightarrow \Phi^{(r)}(\mathbf{x}) \geq 0$ use $\overline{F}$ $y^{(r)} = 0 \Rightarrow \Phi^{(r)}(\mathbf{x}) \leq 0$ use $\underline{F}$
$\underline{y}_0$	$\overline{z}_j^r = 0$ use $\overline{\tilde{D}}A^r$ $\overline{z}_j^r = 1$ use $\underline{\tilde{D}}A^r$	$y^{(r)} = 0 \Rightarrow \Phi^{(r)}(\mathbf{x}) \leq 0$ use $\overline{F}$ $y^{(r)} = 1 \Rightarrow \Phi^{(r)}(\mathbf{x}) \geq 0$ use $\underline{F}$

is used and the difference $\underline{y}^{(r)} - \underline{y}^{(q)}$ for $\partial \underline{y}_0 / \partial F^{(r)}(\mathbf{x})$, respectively.

The proposed solution is a generalisation of the approach described by Nowicki (2008; 2009) who depicted a system for classification, so the answers of the rules can be 0 or 1. A comparison of this approach with ours one is presented in Tab. 2.

4.2.2. Calculation of the rule consequence. The are two models $\mathcal{U}$ and $\mathcal{L}$ for each rule with consequences. The missing values in the data tuple are imputed with maximum and minimum values. Two values are calculated:

$$y_l(\mathbf{x}) = \sum_{a=0}^{A} \min\left(\underline{x}_a \underline{p}_a, \overline{x}_a \underline{p}_a\right), \tag{47}$$

$$y_u(\mathbf{x}) = \sum_{a=0}^{A} \max\left(\underline{x}_a \overline{p}_a, \overline{x}_a \overline{p}_a\right), \tag{48}$$

where $x_0 = 1$, cf. Eqn. (5). Then for the r-th rule we get

$$\overline{y}^{(r)}(\mathbf{x}) = \max\left(y_l^{(r)}(\mathbf{x}), y_u^{(r)}(\mathbf{x})\right), \tag{49}$$

$$\underline{y}^{(r)}(\mathbf{x}) = \min\left(y_l^{(r)}(\mathbf{x}), y_u^{(r)}(\mathbf{x})\right). \tag{50}$$

4.3. Error measure. In the ANNBFIS system the RMSE (Root Mean Square Error) is calculated with the formula (Czogała and Łęski, 2000)

$$E = \sqrt{\frac{1}{X} \sum_{\mathbf{x} \in \mathbb{X}} \left[y_0(\mathbf{x}) - y(\mathbf{x})\right]^2}, \tag{51}$$

where $y_0(\mathbf{x})$ is the answer of the ANNBFIS system for the tuple $\mathbf{x}$, $y(\mathbf{x})$ is the original expected value of the decision attribute of the tuple.

For our system the above formula should be modified. The system elaborates two answers, $\overline{y}_0(\mathbf{x})$ and $\underline{y}_0(\mathbf{x})$, for each tuple $\mathbf{x}$. Instead of one value, our system returns the interval $\left[\underline{y}_0(\mathbf{x}), \overline{y}_0(\mathbf{x})\right]$. The deviation $\Delta y(\mathbf{x})$ of the original value from the returned interval is determined as

$$\Delta y(\mathbf{x}) = \begin{cases} y(\mathbf{x}) - \overline{y}_0(\mathbf{x}), & \text{if } y(\mathbf{x}) > \overline{y}_0(\mathbf{x}), \\ 0, & \text{if } y(\mathbf{x}) \in \left[\underline{y}_0(\mathbf{x}), \overline{y}_0(\mathbf{x})\right], \\ \underline{y}_0(\mathbf{x}) - y(\mathbf{x}), & \text{if } y(\mathbf{x}) < \underline{y}_0(\mathbf{x}). \end{cases} \tag{52}$$

Such a definition of the deviation would promote models that elaborate very wide intervals in the answer. This is why the length of the interval is also taken into account in measuring the system answer:

$$E = \sqrt{\frac{1}{X} \sum_{\mathbf{x} \in \mathbb{X}} \left(\left[\Delta y(\mathbf{x})\right]^2 + \left[\overline{y}_0(\mathbf{x}) - \underline{y}_0(\mathbf{x})\right]^2 \right)}. \tag{53}$$

5. Experiments

The experiments were conducted on real life data.

5.1. Data sets. Two data sets were used in the experiments. Gas Furnace is a popular real life data set depicting the concentration of methane (x) and carbon dioxide (y) in a gas furnace (Box and Jenkins, 1970). The data set contains 290 tuples organised according to the template $[y_{n-1}, \ldots, y_{n-4}, x_{n-1}, \ldots, x_{n-6}, y_n]$.

Table 3. Comparison of ANNBFIS and our system for full data sets without missing values.

data set	RMSE	
	ANNBFIS	our approach
Gas Furnace	0.2285	0.2373
Concrete	16.1127	16.1249

Concrete Compressive Strength is a real life data set describing the parameters of a concrete sample and its strength (Yeh, 1998). The attributes stand for the cement ratio, the amount of blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate, age; the last attribute is the concrete compressive strength. All attributes are numerical.

The data sets were prepared in various manners. The first data set collection comprises a data set with missing 1, 2, 5, 10, 20, 50, 75% of attribute values. In the second collection there are 10% of values of one attribute missing. For Gas Furnace the missing attribute in this collection is the 1-st to 10-th, and for Concrete—1-st to 8-th.

To test the influence of A_t (cf. Section 4.1.1) on the error of the system and time of model creation, one more version of the Gas Furnace data set was prepared. The original data set contains 10 attributes. The prepared data set contains an equal number of tuples with missing 0 to 10 attributes.

The data sets are not normalised.

5.2. Results. The experiments were executed for Knowledge Approximation (KA) where training and test sets are the same. If not stated otherwise, the parameter A_t (Section 4.1.1) is set to 3.

The first experiment was performed to show that the proposed fuzzy rough system reduces to the original ANNBFIS system in the case of a full data set with no missing values. The theoretical analysis (cf. Eqns. (35), (38) and (39)) clearly states that for full data sets the results elaborated by our system and ANNBFIS should be the same. Table 3 presents comparisons of the RMSE of results elaborated by ANNBFIS and by our system. The results are not exactly the same due to probable numerical imprecision. The Table 4 compares the results elaborated for the first 10 tuples from the Gas Furnace data set without missing values. A similar comparison for the last 10 items of the Concrete data set is gathered in Table 5 without missing values.

The next step of experiments was executed for the collection of data sets with missing 1, 2, 5, 10, 20, 50, 75% of values. The results are presented in Tables 6 and 7. The tables do not contain results for data sets with missing 50 and 75% of values. In these sets there are no full tuples with all attributes and the "lower" data set is empty, $\underline{\mathbb{X}} = \varnothing$. The results express the expected feature that the interval width increases with the ratio of missing values in the data set. For the Concrete data set the distance of the expected values from the interval elaborated by the system decreases with the increasing ratio of missing values.

Table 4. Comparison of answers elaborated by ANNBFIS and our system for the Gas Furnace data set without missing values.

expected value	elaborated value	
	ANNBFIS	our approach
52.70	52.75	52.75
52.40	52.32	52.32
52.20	52.12	52.11
52.00	52.07	52.07
52.00	51.93	51.93
52.40	52.15	52.15
53.00	52.92	52.93
54.00	53.76	53.78
54.90	55.05	55.08
56.00	55.92	55.94

Table 5. Comparison of answers elaborated by ANNBFIS and our system for the Concrete data set without missing values.

expected value	elaborated value	
	ANNBFIS	our approach
42.14	36.16	36.10
31.87	27.86	27.81
41.54	34.56	34.52
39.45	35.75	35.72
37.91	34.45	34.42
44.28	40.06	40.04
31.17	33.59	33.54
23.69	26.14	26.12
32.76	28.53	28.51
32.40	32.11	32.09

In the case of data sets with missing values the original data set is preprocessed and two data sets are obtained: the lower $\underline{\mathbb{X}}$ and upper $\overline{\mathbb{X}}$ data sets. Tables 8 and 9 show the numbers of tuples in both preprocessed data sets (X_u for upper and for X_l lower) and execution time (creation of models and elaboration of answers).

The lower set $\underline{\mathbb{X}}$ is the result of marginalisation of tuples with missing data. Based this data set the localisation of the core of the cluster is calculated. Unfortunately, the lack of some values may influence localisation precision. If the data set has more missing values, the lower data set has fewer tuples and the localisation of the cluster's core is less reliable. This feature is illustrated in Fig. 7.

In creation of set $\overline{\mathbb{X}}$ the imputation method is used. This process is described in Section 4.1.1. The crucial parameter is A_t. Table 10 presents the results elaborated for

Table 6. Results elaborated for various percentages of missing values from the Gas Furnace data set.

missing values	error for Gas Furnace		
	RMSE	deviation	interval
full	0.2373	0.2373	0.0000
1%	1.7514	0.2903	1.7271
2%	2.3386	0.2158	2.3287
5%	3.4129	0.3708	3.3927
10%	5.3096	0.8010	5.2488
20%	10.3632	2.6002	10.0316

Table 7. Results elaborated for various percentages of missing values from the Concrete data set.

missing values	error for Concrete		
	RMSE	deviation	interval
full	16.1249	16.1249	0.0000
1%	15.4528	14.9117	4.0536
2%	14.2167	12.9833	5.7920
5%	15.0515	12.1604	8.8696
10%	15.7550	11.1346	11.1463
20%	21.0037	11.4106	17.6338

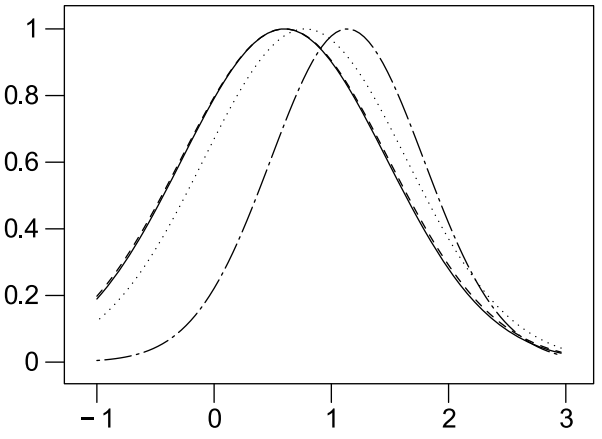

Fig. 7. Influence of missing values on the localisation of the cluster core ('lower' part). The graph shows the fourth attribute of the Gas Furnace data set for a full data set (solid line), a data set with 1% (dashed line), 5% (dotted line) and 20% (dotted and dashed line) missing values.

the Gas Furnace data set in the function of the A_t parameter. Higher values of A_t lead to better values of the RMSE, lower values of deviations and narrower intervals (so the models are better). On the other hand, higher values of A_t lead to larger upper sets $\overline{\mathbb{X}}$ and longer time of execution. The time needs grow quicker than linearly.

The influence of the absence of values in various attributes was analysed with the collection of data sets described in Section 5.1. The average of the squared interval length for missing attributes is presented in Table 11. Different behaviour for data sets used can be observed. The averages of squared interval lengths differ highly for various attributes of the Concrete data set. On the other hand, this phenomenon is not so visible for the Gas Furnace data set. This can be explained with the semantics of attributes. The Gas Furnace data set is a time series data set. The subsequent tuples are shifted. This means that the value of a certain attribute a in the n-th tuple is equal to the values of the $(a+1)$-st attribute in the $(n+1)$-st tuple. That means that for the whole data set no attribute can be labelled with physical meaning. A different situation is in the case of the Concrete data set. In this set each attribute has certain physical meaning. The meanings of attributes of these data sets are listed in Section 5.1. Perhaps this phenomenon can be helpful for evaluation and selection of attributes.

Table 8. Number of tuples in lower (X_l) and upper (X_u) data sets and execution time in seconds for the Gas Furnace data set.

missing values	X_l	X_u	t
full	290	290	5
1%	265	323	8
2%	239	353	10
5%	166	472	13
10%	116	720	25
20%	31	1225	67

Table 9. Number of tuples in lower (X_l) and upper (X_u) data sets and execution time in seconds for the Concrete data set.

missing values	X_l	X_u	t
full	290	290	52
1%	952	1108	50
2%	878	1202	51
5%	708	1496	62
10%	438	2182	110
20%	153	3741	314

The Gas Furnace data set is a time series. Figure 8 presents the original values (black solid line) and the values elaborated by the system (grey line). The model was prepared with a full data set and then tested with the same data. Figures 9 and 10 illustrate the experiment when the models were created with data sets with missing values (10% and 20%, respectively). Both models were tested with a full data set. The black solid line depicts the expected values, the gray regions in both figures is the upper-lower interval for each data tuple. Figures 11 and 12 depict the opposite paradigm: the model is prepared with a full data set and tested with a data set with missing values

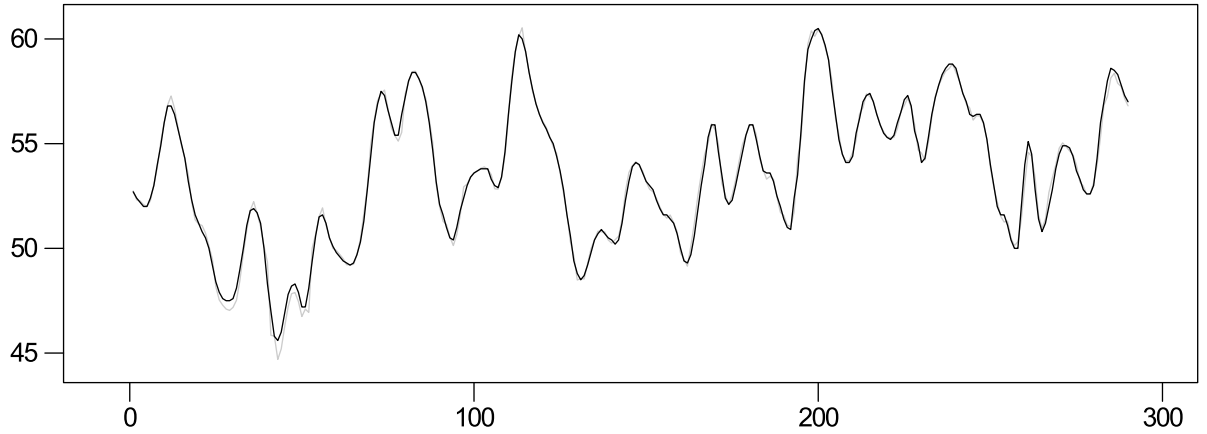

Fig. 8. Results elaborated for the Gas Furnace data set, a model with $c = 3$ rules created with a full data set. The figure presents the expected values (black solid line) and the elaborated results (grey solid line) for the full data set.

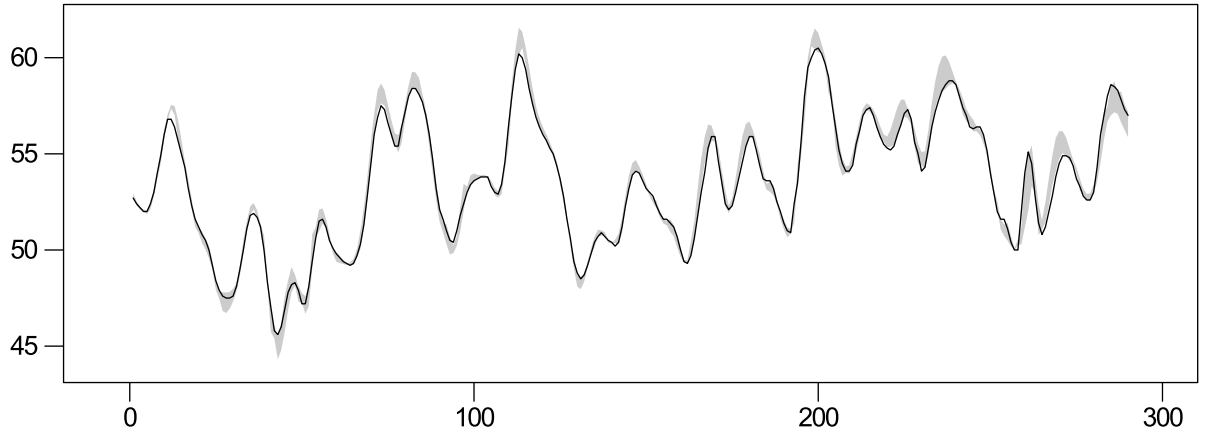

Fig. 9. Results elaborated for the Gas Furnace data set, a model with $c = 3$ rules created with a data set with missing 10% of values. The figure presents the expected values (black solid line) and the elaborated results (grey region—intervals) for the full data set.

Table 10. Influence of the A_t parameter on the precision of the system and the time of model creation. The number X_l of tuples in the lower data set is the same in all situations, $X_l = 29$.

A_t	RMSE	deviations	intervals	X_u	t
1	8.9016	0.8419	8.8617	2639	189
2	8.4191	0.8896	8.3720	2639	190
3	8.6620	0.8151	8.6236	2697	194
4	8.7093	0.8720	8.6656	2929	230
5	8.2022	0.7798	8.1650	3567	357
6	9.2521	0.6643	9.2282	5075	687
7	7.8133	0.5839	7.7914	8381	1924

(10% and 20%, respectively).

Figures 13 and 14 present the results elaborated for the Gas Furnace data set with 1% and 10% missing values respectively. The model was created and tested with the same data set (data approximation paradigm). For a data set with 10% missing values, the upper-lower region is wider than for a 1% missing value data set.

Table 11. Influence of the missing attribute on the interval width. The given value is the average of a squared interval length.

attr.	Gas Furnace	Concrete
1st		4.4391
2nd	0.0733	2.6458
3rd	0.0475	1.6867
4th	0.0689	1.8345
5th	0.0201	0.7512
6th	0.0662	0.4489
7th	0.0727	0.6355
8th	0.0587	9.7130
9th	0.0435	–
10th	0.0497	–

The above mentioned figures show that the proposed system better handles creating models based on missing values and then elaborating the answer for full tuples. The

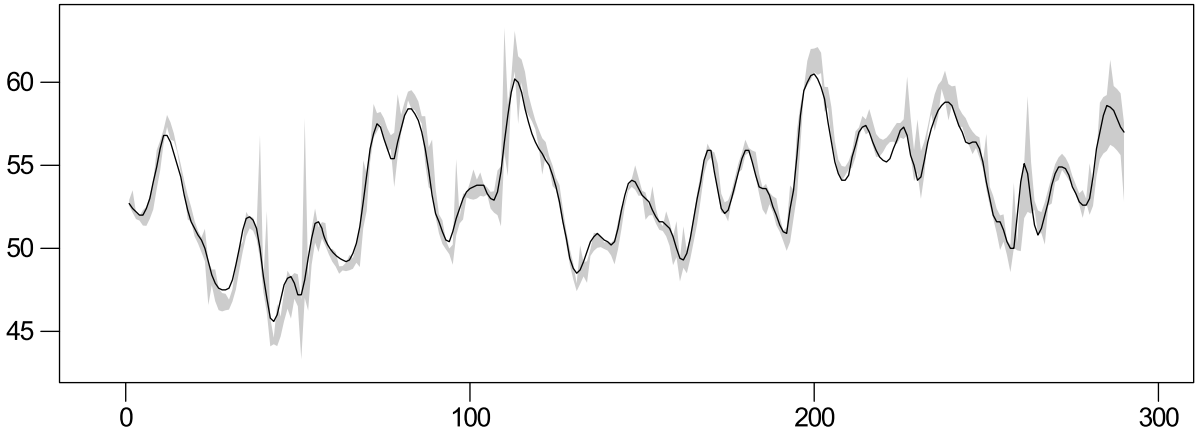

Fig. 10. Results elaborated for the Gas Furnace data set, a model with $c = 3$ rules created with a data set with missing 20% of values. The figure presents the expected values (black solid line) and the elaborated results (grey region—intervals) for the full data set.

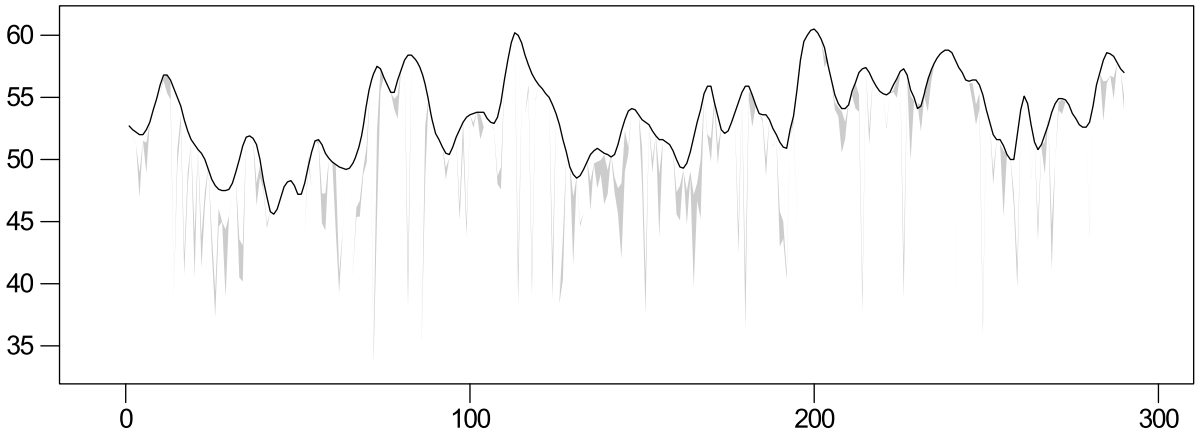

Fig. 11. Results elaborated for the Gas Furnace data set, a model with $c = 3$ rules created with a full data set. The figure presents the expected values (black solid line) and the elaborated results (grey region—intervals) for the data set with missing 10% of values.

opposite paradigm—creation of the model based on complete data and then elaboration of answers for incomplete tuples—results in poorer results. This can be easily seen when comparing Figs. 9 and 11. The results reveal an important feature of the proposed system. The models created on data sets with missing values elaborate the upper and lower answer values that mostly embrace the expected value. The more values miss from the train data set, the wider the lower-upper interval of the answer. This is expected behaviour. Unfortunately, when the model (independently, whether created with a full data set or a data set with missing values) is tested with a data set with missing value, the results are less advantageous. The intervals are not very wide, but as a whole they are often far from expected values. Mostly the whole upper-lower interval is shifted towards lower values, so the expected value exceeds the upper boundary of the interval elaborated by the system.

In the work of Nowicki (2010) a system with rough answers is applied for classification. The expected answers are $\{0, 1\}$. If both upper and lower answers are greater, then half the tuple is classified to the class labelled with 1. If both answers are less than a half, the tuple is labelled with zero. If one answer is greater and the other less than a half, the system gives no answer. Also in our system the double answer for each given tuples is elaborated, but we restrain from a decision whether the rough answer is precise enough or not. This decision is left to the user of the system. Sometimes the interval between upper and lower answers is big, but maybe the user wishes to take such an answer into account.

6. Conclusions

The paper presents a Neuro-Fuzzy System (NFS) based on the ANNBFIS system for data sets with missing values. The type of data is difficult to handle with the NFS. The described system is a complete one. This means that is

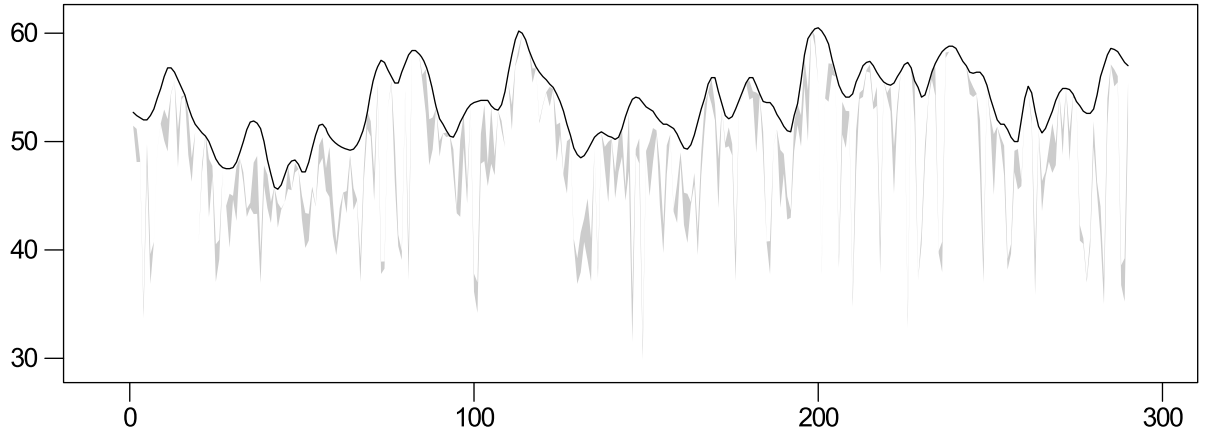

Fig. 12. Results elaborated for the Gas Furnace data set, a model with $c = 3$ rules created with a full data set. The figure presents the expected values (black solid line) and the elaborated results (grey region—intervals) for the data set with missing 20% of values.

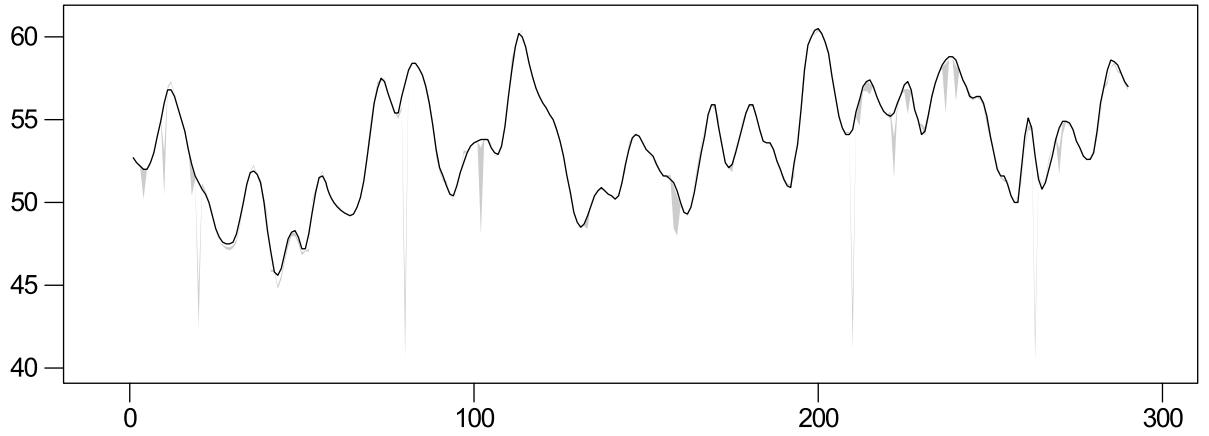

Fig. 13. Results elaborated for the Gas Furnace data set, a model with $c = 3$ rules created with a data set with missing 1% of values. The figure presents the expected values (black solid line) and the elaborated results (grey region—intervals) for the data set with missing 1% of values.

able to create the fuzzy model (rule base) based on missing value data set and then elaborate answers for missing value tuples. The missing data are preprocessed with two most often used methods for handling missing values (marginalisation and imputation).

The system joins fuzzy and rough set theories. The rules in the rule base are extracted with a special modified clustering algorithm. This algorithm creates rough fuzzy clusters.

The system can handle both full and missing value tuples. If the system was created with full data and elaborates answers for full values tuples, it is theoretically and practically equal to its parent system ANNBFIS. The experiments confirm this feature.

If the ratio of missing values in test tuples grows, the deviation of the answer from the expected value remains more or less the same. But the width of the returned interval grows. This is an expected behaviour. The more missing values in the data set, the more rough the answers. The proposed system better handles the situation when the model is created with a data set with missing values and then elaborates the answers for full tuples. Creation of a model on missing value data sets gives poorer results.

Acknowledgment

This work was supported by the European Union within the European Social Fund (grant agreement no. UDA-POKL.04.01.01-00-106/09).

The author is grateful to the anonymous referees for their constructive comments that helped him to improve the paper.

References

Acuña, E. and Rodriguez, C. (2004). The treatment of missing values and its effect in the classifier accuracy, *in* D. Banks,

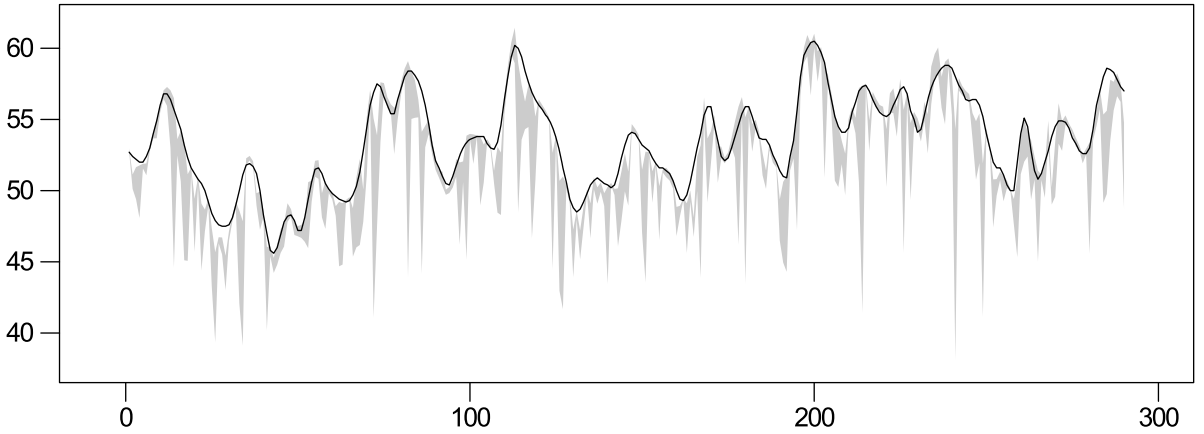

Fig. 14. Results elaborated for the Gas Furnace data set, model with $c = 3$ rules created with a data set with missing 10% of values. The figure presents the expected values (black solid line) and the elaborated results (grey region—intervals) for the data set with missing 10% of values.

L. House, F. McMorris, P. Arabie and W. Gaul (Eds.), *Classification, Clustering and Data Mining Applications,* Springer, Berlin/Heidelberg, pp. 639–648.

Box, G.E.P. and Jenkins, G. (1970). *Time Series Analysis, Forecasting and Control*, Holden-Day, Oakland, CA.

Chan, L.S., Gilman, J.A. and Dunn, O.J. (1976). Alternative approaches to missing values in discriminant analysis, *Journal of the American Statistical Association* **71**(356): 842–844.

Cooke, M., Green, P., Josifovski, L. and Vizinho, A. (2001). Robust automatic speech recognition with missing and unreliable acoustic data, *Speech Communication* **34**: 267–285.

Czogała, E. and Łęski, J. (2000). *Fuzzy and Neuro-Fuzzy Intelligent Systems*, Series in Fuzziness and Soft Computing, Physica-Verlag, Heidelberg/New York, NY.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B* **39**(1): 1–38.

Dubois, D. and Prade, H. (1990). Rough fuzzy sets and fuzzy rough sets, *International Journal of General Systems* **17**(2): 191–209.

Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact, well separated clusters, *Journal Cybernetics* **3**(3): 32–57.

Farhangfar, A., Kurgan, L. and Dy, J. (2008). Impact of imputation of missing values on classification error for discrete data, *Pattern Recognition* **41**(12): 3692–3705.

Farhangfar, A., Kurgan, L. and Pedrycz, W. (2007). A novel framework for imputation of missing values in databases, *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans* **37**(5): 692–709.

Fuller, W.A. and Kim, J.K. (2005). Hot deck imputation for the response model, *Survey Methodology* **31**(2): 139–149.

Ghahramani, Z. and Jordan, M. (1995). Learning from incomplete data, *Technical report*, Lab Memo No. 1509, CBCL Paper No. 108, MIT AI Lab, Cambridge, MA.

Grzymala-Busse, J. (2006). A rough set approach to data with missing attribute values, *in* G. Wang, J. Peters, A. Skowron and Y. Yao (Eds.), *Rough Sets and Knowledge Technology*, Lecture Notes in Computer Science, Vol. 4062, Springer, Berlin/Heidelberg, pp. 58–67.

Grzymala-Busse, J.W. and Hu, M. (2001). A comparison of several approaches to missing attribute values in data mining, *in* W. Ziarko and Y. Yao (Eds.), *Rough Sets and Current Trends in Computing*, Lecture Notes in Computer Science, Vol. 2005, Springer, Berlin/Heidelberg, pp. 378–385.

Hathaway, R. and Bezdek, J. (2001). Fuzzy c-means clustering of incomplete data, *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* **31**(5): 735–744.

Himmelspach, L. and Conrad, S. (2010). Fuzzy clustering of incomplete data based on cluster dispersion, *in* E. Hüllermeier, R. Kruse and F. Hoffmann (Eds.), *Computational Intelligence for Knowledge-Based Systems Design, 13th International Conference on Information Processing and Management of Uncertainty, IPMU 2010, Dortmund, Germany, June 28–July 2, 2010. Proceedings*, Lecture Notes in Computer Science, Vol. 6178, Springer, Berlin/Heidelberg, pp. 59–68.

Hwang, C. and Rhee, F.C.-H. (2004). An interval type-2 fuzzy c spherical shells algorithm, *Proceedings of the 2004 IEEE International Conference on Fuzzy Systems, Budapest, Hungary*, pp. 1117–1122.

Korytkowski, M., Nowicki, R., Scherer, R. and Rutkowski, L. (2008). Ensemble of rough-neuro-fuzzy systems for classification with missing features, *IEEE International Conference on Fuzzy Systems, FUZZ-IEEE (IEEE World Congress on Computational Intelligence), Hong Kong, China*, pp. 1745–1750.

Lakshminarayan, K., Harp, S.A. and Samad, T. (1999). Imputation of missing data in industrial databases, *Applied Intelligence* **11**(3): 259–275, DOI: 10.1023/A:1008334909089.

Łęski, J. (2008). *Neuro-Fuzzy Systems*, Wydawnictwa Naukowo-Techniczne, Warsaw, (in Polish).

Łęski, J. and Czogała, E. (1999). A new artificial neural network based fuzzy inference system with moving consequents in if-then rules and selected applications, *Fuzzy Sets and Systems* **108**(3): 289–297.

Mamdani, E.H. and Assilian, S. (1975). An experiment in linguistic synthesis with a fuzzy logic controller, *International Journal of Man-Machine Studies* **7**(1): 1–13.

Nowicki, R. (2006). Rough-neuro-fuzzy system with MICOG defuzzification, *2006 IEEE International Conference on Fuzzy Systems, Vancouver, Canada*, pp. 1958–1965.

Nowicki, R. (2008). On combining neuro-fuzzy architectures with the rough set theory to solve classification problems with incomplete data, *IEEE Transactions on Knowledge and Data Engineering* **20**(9): 1239–1253.

Nowicki, R.K. (2009). Rough-neuro-fuzzy structures for classification with missing data, *IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics* **39**(6): 1334–1347.

Nowicki, R.K. (2010). On classification with missing data using rough-neuro-fuzzy systems, *International Journal of Applied Mathematics and Computer Science* **20**(1): 55–67, DOI: 10.2478/v10006-010-0004-8.

Pawlak, Z. (1982). Rough sets, *International Journal of Parallel Programming* **11**(5): 341–356.

Pedrycz, W. (1998). Conditional fuzzy clustering in the design of radial basis function neural networks, *IEEE Transactions on Neural Networks* **9**(4): 601–612.

Renz, C., Rajapakse, J.C., Razvi, K. and Liang, S.K.C. (2002). Ovarian cancer classification with missing data, *Proceedings of the 9th International Conference on Neural Information Processing, ICONIP'02, Singapore*, Vol. 2, pp. 809–813.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York, NY.

Sugeno, M. and Kang, G.T. (1988). Structure identification of fuzzy model, *Fuzzy Sets and Systems* **28**(1): 15–33.

Takagi, T. and Sugeno, M. (1985). Fuzzy identification of systems and its application to modeling and control, *IEEE Transactions on Systems, Man and Cybernetics* **15**(1): 116–132.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R.B. (2001). Missing value estimation methods for DNA microarrays, *Bioinformatics* **17**(6): 520–525.

Wagstaff, K. (2004). Clustering with missing values: No imputation required, *in* D. Banks, L. House, F.R. Mcmorris, P. Arabie and W. Gaul (Eds.), *Classification, Clustering, and Data Mining Applications (Proceedings of the Meeting of the International Federation of Classification Societies)*, Springer, Berlin/Heidelberg, pp. 649–658.

Wagstaff, K.L. and Laidler, V.G. (2005). Making the most of missing values: Object clustering with partial data in astronomy, *Proceedings of Astronomical Data Analysis Software and Systems XIV, Pasadena, CA*, Vol. 347, pp. 172–176.

Yeh, I. C. (1998). Modeling of strength of high-performance concrete using artificial neural networks, *Cement and Concrete Research* **28**(12): 1797–1808.

Zhang, C., Zhu, X., Zhang, J., Qin, Y. and Zhang, S. (2007). GBKII: An imputation method for missing values, *Advances in Knowledge Discovery and Data Mining* **4426**: 1080–1087.

Zhang, S. (2011). Shell-neighbor method and its application in missing data imputation, *Applied Intelligence* **35**(1): 1–11, DOI: 10.1007/s10489-009-0207-6.

Krzysztof Simiński received the M.Sc. and Ph.D. degrees in computer science from the Silesian University of Technology (Gliwice, Poland) in 2006 and 2009, respectively. His main interests include data mining, fuzzy reasoning, and natural language processing.

Received: 1 February 2011
Revised: 11 July 2011
Re-revised: 25 October 2011