Computer Science • 15 (4) 2014 http://dx.doi.org/10.7494/csci.2014.15.4.441

Marcin Sikora
Krzysztof Simiński

COMPARISON OF INCOMPLETE DATA HANDLING TECHNIQUES FOR NEURO-FUZZY SYSTEMS

Abstract *Real-life data sets sometimes miss some values. The incomplete data needs specialized algorithms or preprocessing that allows the use of the algorithms for complete data. The paper presents a comparison of various techniques for handling incomplete data in the neuro-fuzzy system ANNBFIS. The crucial procedure in the creation of a fuzzy model for the neuro-fuzzy system is the partition of the input domain. The most popular approach (also used in the ANNBFIS) is clustering. The analyzed approaches for clustering incomplete data are: preprocessing (marginalization and imputation) and specialized clustering algorithms (PDS, IFCM, OCS, NPS). The objective of our research is the comparison of the preprocessing techniques and specialized clustering algorithms to find the the most-advantageous technique for handling incomplete data with a neuro-fuzzy system. This approach is also the indirect validation of clustering.*

Keywords incomplete data, marginalization, imputation, neuro-fuzzy system, ANNBFIS, PDS, IFCM, OCS, NPS

1. Introduction

Real-life data sets sometimes miss certain values. This may be the result of an improper acquisition of data, the failure of detectors, the reluctance to answer questions in a questionnaire, etc.

This incomplete data needs specialized algorithms or preprocessing that allows the use of the algorithms for complete data. The preprocessing techniques enumerate two classes:

1. marginalization – deletion of incomplete data items or deletion of incomplete attributes;
2. imputation – substitution of missing values with some values (constant, average, median, etc.).

In this paper, we focus on the analysis of incomplete data with neuro-fuzzy systems. The core of the neuro-fuzzy system is the fuzzy-rule base composed of fuzzy implications. There are three main techniques of automatic creation of the fuzzy-rule base for neuro-fuzzy systems. These are:

- grid partition – the oldest approach, the input domain is split into hyperrectangular grid, vulnerable to number of dimensions [12];
- scatter partition (clustering) – the most popular technique, may leave some regions uncovered by any rule [6]
- hierarchical partition – the latest method, has advantages of grid and scatter partitions but avoids their faults [19, 20, 25, 26].

The objective of our research is the comparison of the preprocessing techniques and specialized clustering algorithms to find the most-advantageous technique for handling incomplete data with a neuro-fuzzy system. In our experiments, we use the ANNBFIS neuro-fuzzy system [6]. This system uses clustering for partitioning input domain. We analyze the generalization ability of the systems created with various techniques of incomplete-data handling. This approach is also the indirect validation of clustering, because the partition of the domain is crucial for neuro-fuzzy systems. Many indices of clustering quality exist [35, 21, 2, 6, 17], but they do not always reflect practical applications. Our approach can be treated as some practical indirect validation of clustering algorithms.

The paper is organized in the following way: Sec. 2 discusses the methods of incomplete data handling (preprocessing and specialized algorithms). Sec. 3 shortly introduces the ANNBFIS neuro-fuzzy system. Sec. 4 describes the experiments and their results. And finally Sec. 5 sums up the paper.

2. Clustering of incomplete data

Clustering of incomplete data can be divided into two paradigms:

- preprocessing of incomplete data and application of algorithms for complete data,
- specialized algorithms for incomplete data.

2.1. Preprocessing

Preprocessing is a common technique for handling incomplete data with algorithms that have proven to be efficient for the complete data [10]. There are two essential approaches: (1) marginalization of incomplete data items, (2) imputation of missing values with several techniques.

Both marginalization and imputation are often used due to their simplicity. Preprocessing distorts the data: marginalization loses some information, but imputation (on the other hand) may add non-existing or meaningless information [34].

2.1.1. Marginalization

Marginalization is the simplest method of treating incomplete data. This method deletes incomplete objects from the data set [33, 11] or deletes incomplete attributes from the set of attributes [4]. The latter approach reduces the dimensionality of the data set. Marginalization may severely reduce the amount of data and deletes the potential knowledge in the incomplete data.

2.1.2. Imputation

The missing values can be imputed with various values. The simplest way is imputation with constant values, zeros, random values, mean values (over all data set [18], over the class of the data item [9]). The more-sophisticated techniques are nearest-neighbor selection [36, 37], Expectation-Maximization (EM) algorithm [8] or hot-deck [23] and cold-deck [13] techniques to avoid imputation of non-existing values.

Imputation with the average of existing values of the attribute is commonly used. This method is very simple, but has some disadvantages. The average value is vulnerable to outliers and may have no sense [34], or may present non-existing values [1]. Imputation with median values is less vulnerable to outliers and imputes the missing values with existing values. The mean time of calculation of average and median is a linear function of data size. A popular method is imputation with values elaborated on k nearest neighbors (kNN). Commonly average and median are used. The disadvantage of this method is high cost of the selection of k nearest neighbors from all data items and choosing the most advantageous value of k parameter. The imputed values are not labeled and cannot be distinguished from the original data. The conclusions drawn from imputed data are not always reliable enough [33].

The marginalization, imputation, and rough sets are combined and used to preserve the distinction between original and imputed data items [28, 29].

2.2. Specialized clustering algorithms

Many specialized algorithms base on the fuzzy c-mean (FCM) algorithm [7]. The FCM clustering minimizes the objective function

$$J(\mathbf{U}, \mathbf{V}) = \sum_{c=1}^{C} \sum_{k=1}^{K} (u_{ck})^m \left\| \mathbf{x}_k - \mathbf{v}_c \right\|^2 \tag{1}$$

with constraints

$$\forall k \in \mathbb{K} : \sum_{c=1}^{C} u_{ck}, \tag{2}$$

where C stands for number of clusters, K – number of objects (vectors), $\mathbf{x} = [x_1, x_2, \ldots, x_D]^{\mathrm{T}}$ – object (data vector) with D attributes, $\mathbf{v} = [v_1, v_2, \ldots, v_D]^{\mathrm{T}}$ – cluster center and u_{ck} is membership values of the k-th object to the c-th cluster. The values u_{ck} constitute the matrix $\mathbf{U}$ with C rows and K columns. The parameter m commonly equals 2. The center of the c-th cluster is elaborated with the formula

$$\forall c \in [1, C] : \mathbf{v}_c = \frac{\sum_{k=1}^{K} (u_{ck})^m \mathbf{x}_k}{\sum_{k=1}^{K} (u_{ck})^m} \tag{3}$$

The clusters' centers are gathered into matrix $\mathbf{V} = \left[\mathbf{v}_1^{\mathrm{T}}, \mathbf{v}_2^{\mathrm{T}}, \ldots, \mathbf{v}_C^{\mathrm{T}}\right]^{\mathrm{T}}$, thus the matrix $\mathbf{V}$ has C rows, each of them representing the center of one cluster (with D attributes).

The algorithm proposed in [28, 27, 29, 30] applies rough sets to the clustering of data with missing values. This algorithm will not be further analyzed in our paper as it elaborates rough clusters, and ANNBFIS is not able to handle rough clusters.

2.2.1. Partial Distance Strategy

Partial Distance Strategy (PDS) [32, 11, 31] clustering algorithm uses the partial distance. This approach is similar to FCM algorithm [7] and has two differences:

- The attribute d of center of cth cluster is calculated with formula

$$v_{cd} = \frac{\sum_{i=1}^{X} (u_{ci})^m z_{id} x_{id}}{\sum_{i=1}^{X} (u_{ci})^m z_{id}}, \tag{4}$$

  where $z_{id} \in \{0, 1\}$ denotes whether the d-th attribute of i-th data item exists (1) or not (0).

- The distance t of ith data item from cluster c is calculated with formula:

$$t_{ci} = \frac{D}{\sum_{d=1}^{D} z_{id}} \sum_{d=1}^{D} (x_{id} - v_{cd})^2 z_{id} \tag{5}$$

2.2.2. Optimal Completion Strategy

The Optimal Completion Strategy (OCS) imputes the missing values in each iteration of clustering algorithm. The missing values are initially imputed with random values. After the cluster centers and membership values are calculated, the missing values are imputed in each iteration with formula

$$\hat{x}_{id} = \frac{\sum_{c=1}^{C} (u_{ci})^m v_{cd}}{\sum_{c=1}^{C} (u_{ci})^m}. \tag{6}$$

2.2.3. Improved Fuzzy C-Means

The improved fuzzy c-means (IFCM) [32] imputes the missing values iteratively. The clusters are elaborated with full data examples, then the missing values are imputed with weighted mean of values of missing attributes from elaborated clusters' centers:

$$\hat{x}_{id} = \frac{\sum_{c=1}^{C} u_{ci} v_{cd}}{\sum_{c=1}^{C} u_{ci}}. \tag{7}$$

The weights are the membership values of the object in question to the found clusters.

2.2.4. Nearest Prototype Strategy

The Nearest Prototype Strategy (NPS) [11] is similar to the OCS and IFCM methods. The missing values in an object k are recalculated in each iteration, instead of applying the formulae (6) or (7) respectively, the nearest prototype (object with all attributes) is found and the missing values in object k are substituted with respective values of the nearest prototype.

3. ANNBFIS neuro-fuzzy system

For the experiments we use the ANNBFIS neuro-fuzzy system [6]. We will not describe in detail the ANNBFIS system and we will limit ourselves only to highlighting some essential features. The ANNBFIS system is based on fuzzy rules. Each rules represents the fuzzy implication. The premises of the rules are constituted by the fuzzy set with Gaussian membership function. The consequences are formed with isosceles triangles. The localization of the triangles is a function of the input parameters. This is why the system is also called a *neuro-fuzzy system with moving (parametrized) consequences* [15]. The system implements the logical interpretation of the fuzzy rules. Thus the value of each rule is a fuzzy set. These sets are then aggregated into answer of the system and defuzzified into crisp output.

The premises of the fuzzy rules are elaborated from the clustering of the train data. In the original ANNBFIS the FCM algorithm [7] is used. The fuzzy rules are represented by the fuzzy implication. In the ANNBFIS the Reichenbach implication [22] is used although it is possible to use other implication (for details see [6]).

The ANNBFIS system is provided with mechanisms for modifying of the parameters to better fit the presented data. The parameters of the premises of the rules are tuned with the gradient method. The parameters of the rule consequences are also tuned with the gradient method with one exception: the linear coefficients for the calculation of the localization of the consequence sets are calculated with the pseudoinverse matrix.

4. Experiments

The experiments were executed on three data sets:

- 'Gas furnace' is a popular data set describing the carbon dioxide concentration in fumes of the gas furnace [3]. The downloaded[1,2] data is the base for 290 tuples created with the template [6, 5, 14]:

  $$[y(n-1), \ldots, y(n-4), x(n-1), \ldots, x(n-6), y(n)]. \tag{8}$$

  The data set was split into disjunctive train and test sets with 145 data items each.
- 'Carbon dioxide concentration' dataset contains real life measurements of some air parameters in a pump deep shaft in one of the Polish coal mines [24]. The parameters (measured in 1 minute intervals) are: CO_2 – concentration of carbon dioxide, Ps – atmospheric pressure, RHOs – relative humidity of the air in the shaft, RHPs – relative humidity of the air near the pump, TOs – air temperature. The dynamic attributes (10-minute sums of the measurements: DCO_2, DPs, DRHOs, DRHPs, DTOs) are added to the tuples. The task is to predict the concentration of the carbon dioxide in 10 minutes. The data are divided into disjunctive train set (700 tuples) and test set (350 tuples) .
- 'Concentration of leukocytes' in blood is modeled with Mackey-Glass equation [16]:

  $$\frac{dx(t)}{dt} = \frac{ax(t-\tau)}{1 + (x(t-\tau))^{10}} - bx(t), \tag{9}$$

  where x is concentration of leukocytes, $a = 0.2$, $b = 0.1$ and $\tau = 17$ are constants. The equation was solved with condition $x(0) = 0.1$ with Runge-Kutt method with step $k = 0.1$ [14]. The data series was the base for creation of tuples with the template

  $$[x(t), x(y-6), x(t-12), x(t-18), x(k+6)]\,. \tag{10}$$

  The data were split into disjunctive train (200 tuples) and test (300 tuples) sets.

We tested two approaches: (1) preprocessing with subsequent FCM clustering algorithm and (2) specialized clustering algorithms. In preprocessing, we used two techniques: object marginalization and imputation (average, median, kNN average, kNN median, constant). The preprocessed (complete) data is passed to the ANNBFIS system. In the second approach, the FCM clustering algorithm is substituted with specialized algorithms (PDS, IFCM, NPS and OCS) in the ANNBFIS. In both approaches, the ANNBFIS was started with four fuzzy rules, number of clustering iterations: 100, number of tuning iterations: 400. Each experiment was repeated 8 times. The data sets were split into disjunctive train and test sets. The train sets lacked 0% (complete data), 1%, 2%, 5%, 10%, 20%, and 40% of values. The data miss

[1] `http://neural.cs.nthu.edu.tw/jang/benchmark/`

[2] `http://www.stat.wisc.edu/~reinsel/bjr-data/`

values completely at random. The test sets were complete and lacked no values. The error E of the system was calculated as root mean square error (RMSE) defined as

$$E(\mathbb{X}) = \sqrt{\sum_{i=1}^{X} \left(y_0(\mathbf{x}_i) - y(\mathbf{x}_i)\right)^2}, \tag{11}$$

where $\mathbb{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_X\}$ stands for the set of data tuples (vectors), $X = |\mathbb{X}|$ is the number of data tuples in the data set, $y_0(\mathbf{x}_i)$ is the answer of the system for the $\mathbf{x}_i$ data vector, $y(\mathbf{x}_i)$ is the desired answer for this vector.

4.1. Results

The root mean square error (RMSE) elaborated with various techniques of handling of missing values for three test data sets are presented in the Tables 1, 2, and 3 and in the Figures 1, 2, and 3.

The tables and figures lack the results for marginalization of the data set with 40% of missing values, because after the marginalization there was not enough data to create the fuzzy model for the neuro-fuzzy system.

The marginalization is the simplest method for handling missing values in data mining. It cannot be used for high ratios of missing values (in our experiments: 40%). The results for marginalization are quite various: for 20% of missing values in the 'Concentration of leukocytes' data set this method achieves the best results of all tested methods (the error is one order of magnitude lower than in other methods), for the 'Gas furnace' data set the results are mediocre, for the 'Carbon dioxide concentration' data set, the RMSE is very high and is not presented in Figure 2 as to not obfuscate the results of other techniques. With the increase in the ratio of missing values, the number of full-data items decreases, the marginalized data set shrink, and the time of calculations reduces. This is clearly observable in Figures 4 and 5.

The imputation of missing values with the constant value (in our experiments: zero) is unstable: it can result in low errors for one data set and high for other.

For all three data sets, the errors achieved for imputation with the kNN median and kNN average ($k = 5$) are among the smallest for 10% and 20% of missing values (what is in concordance with clustering indices [17]). In the case of 40% of missing values, the results are similar with exception for the 'Concentration of leukocytes' data set. For this data set, the imputation with the kNN median elaborates the smallest error, whereas the imputation with kNN average gives the highest errors (very close to the error elaborated with the average imputation). The time of calculation for the kNN imputation techniques is quite long. It is caused by the selection of k nearest neighbors from all data items in the data set. Table 6 and Figure 6 present the influence of the k parameter in kNN imputation. For missing ratio less than 20%, the difference in RMSE for $k = 3$ and $k = 5$ is small. For 40% missing ratio, the errors elaborated for $k = 3$ are significantly lower than for $k = 5$. In this case there is a high number of incomplete data items and increase of k parameter results in wider neighborhoods.

The average or median value from such a wide neighborhood may not substitute the missing value properly and leads to poorer results than in narrower neighborhoods ($k = 3$). The increase in ratio of missing values leads to a quicker-than-linear increase of error.

Table 1

Root mean square error (RMSE) forthe 'Gas furnace' data set ($k = 5$).

method	0%	1%	2%	5%	10%	20%	40%
marginalization	0.5077	0.5400	0.5077	0.5644	0.9971	0.9248	–
average	0.5077	0.5081	0.5085	0.6069	0.7094	0.9971	1.5020
median	0.5077	0.5238	0.5510	0.6251	0.7391	1.0509	1.4707
zeros	0.5077	0.4359	0.4870	0.3549	0.3411	0.3515	0.9284
k-NN average	0.5077	0.5128	0.4601	0.6586	0.5260	0.5426	1.3133
k-NN median	0.5077	0.4617	0.4387	0.6895	0.5313	0.4519	1.3170
PDS	0.5077	0.4939	0.4426	0.3895	0.7317	1.3045	1.4112
OCS	0.5077	0.4427	0.4859	0.6902	0.5682	1.1144	1.4077
IFCM	0.5077	0.4532	0.4911	0.6901	0.5910	1.1433	1.5832
NPS	0.5077	0.4848	0.5573	0.8052	1.0624	1.0111	1.5192

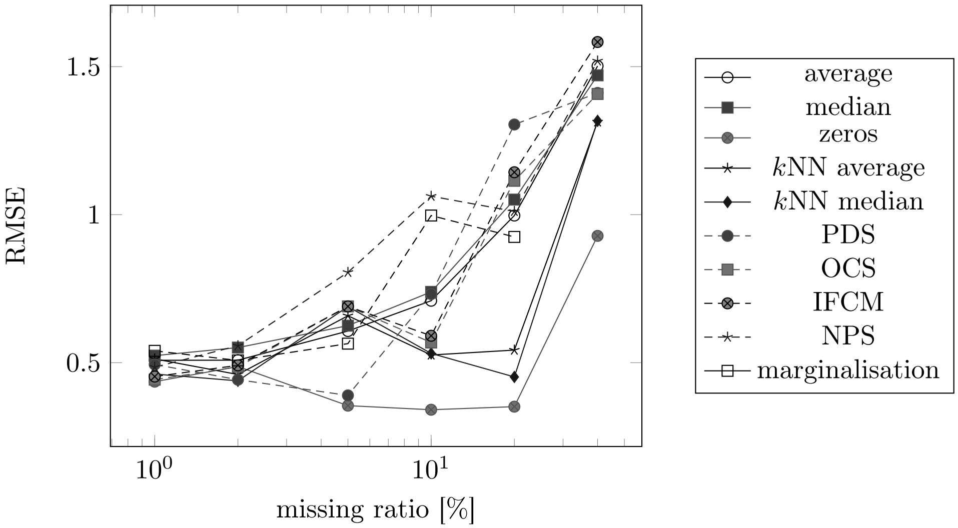

Figure 1. Root mean square error (RMSE) for the 'Gas furnace' data set ($k = 5$).

The specialized clustering algorithms do not achieve low values of errors. The NPS algorithm elaborated high error for the 'Gas furnace' and 'Carbon dioxide concentration' data sets. The NPS algorithm consumes a long time to calculate the results. The specialized algorithms are advantageous in clustering of incomplete data with a high ratio of missing values (> 40%), when clustering results of the specialized algorithms are evaluated with clustering indices [17]. There are many known clustering indices that are not always coherent. Here in our experiments, the clustering quality was tested in an indirect way. The results show that, for high missing ratios (> 40%), the specialized clustering algorithms are not the class with the highest efficacy.

Table 2

Root mean square error (RMSE) for the 'Carbon dioxide concentration' data set ($k = 5$).

method	0%	1%	2%	5%	10%	20%	40%
marginalization	0.3423	0.3517	0.3444	0.4417	0.3779	4.5532	–
average	0.3423	0.3389	0.4935	0.4377	0.4994	0.5230	0.7402
median	0.3423	0.3726	0.5046	0.4403	0.4880	0.5322	0.7222
zeros	0.3423	0.4903	0.4873	0.5554	0.5199	0.6066	0.7042
kNN average	0.3423	0.3310	0.3235	0.3380	0.3313	0.3964	0.7120
kNN median	0.3423	0.3255	0.3484	0.3264	0.3332	0.3655	0.7709
PDS	0.3423	0.3799	0.4801	0.3837	0.4765	0.4862	0.7390
OCS	0.3423	0.3244	0.3731	0.3529	0.4146	0.4384	0.7119
IFCM	0.3423	0.3191	0.4236	0.3750	0.3720	0.4564	0.8622
NPS	0.3423	0.4096	0.4591	0.4372	0.4677	0.4636	1.0295

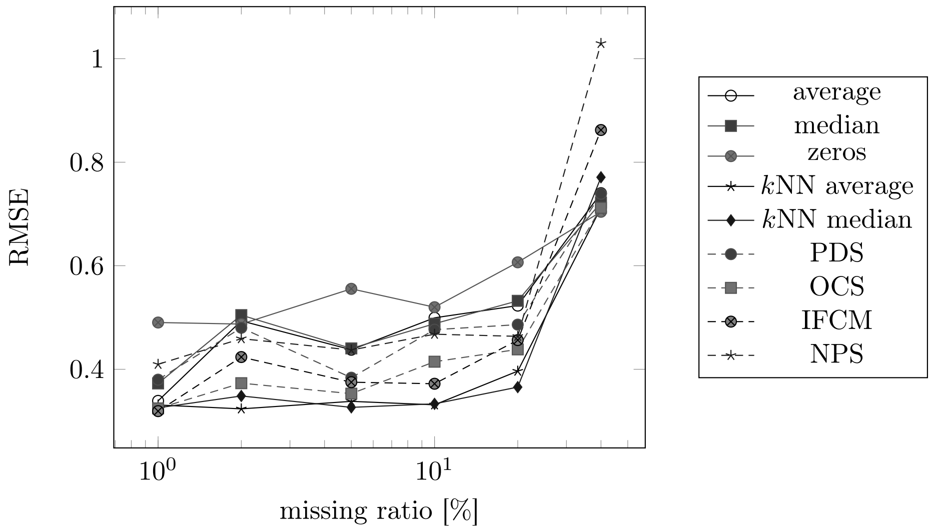

Figure 2. Root mean square error (RMSE) for the 'Carbon dioxide concentration' data set ($k = 5$).

Table 3

Root mean square error (RMSE) for the 'Concentration of leukocytes' data set ($k = 5$).

method	0%	1%	2%	5%	10%	20%	40%
marginalization	0.00053	0.00075	0.00081	0.00063	0.00066	0.00406	–
average	0.00053	0.00319	0.00981	0.01201	0.02395	0.03932	0.10479
median	0.00053	0.00284	0.00905	0.01813	0.02680	0.03602	0.09066
zeros	0.00053	0.00174	0.00409	0.00919	0.02729	0.07626	0.06924
kNN average	0.00053	0.00364	0.00298	0.00428	0.00960	0.02970	0.10611
kNN median	0.00053	0.00169	0.00392	0.00565	0.01225	0.02535	0.04592
PDS	0.00053	0.00167	0.00778	0.03189	0.02629	0.03194	0.06406
OCS	0.00053	0.00253	0.00329	0.00709	0.01816	0.04894	0.08359
IFCM	0.00053	0.00312	0.00347	0.00702	0.01966	0.04596	0.09014
NPS	0.00053	0.00168	0.00397	0.01822	0.02458	0.03189	0.07335

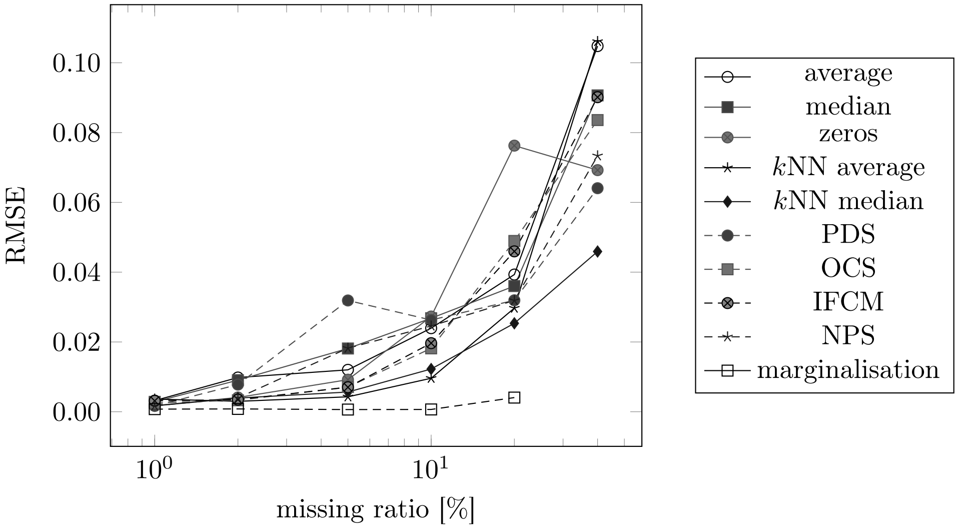

Figure 3. Root mean square error (RMSE) for the 'Concentration of leukocytes' data set ($k = 5$).

Table 4

Time (in [s]) of calculation for the 'Gas furnace' data set ($k = 5$).

method	0%	1%	2%	5%	10%	20%	40%
marginalization	7.013	5.969	5.596	4.208	2.810	1.818	–
average	7.153	6.574	6.585	6.604	6.527	6.545	7.208
median	7.297	6.652	6.521	6.590	6.472	6.469	7.315
zeros	6.560	6.696	6.770	6.754	6.462	7.805	7.167
kNN average	6.593	6.577	6.522	6.609	6.453	6.722	17.832
kNN median	6.504	6.515	6.587	6.643	6.538	6.767	17.845
PDS	6.620	6.845	6.630	6.947	6.883	7.026	7.442
OCS	6.678	6.794	6.725	6.904	7.050	7.104	8.994
IFCM	6.538	7.031	6.904	6.991	7.139	7.818	9.367
NPS	7.511	8.979	9.882	11.639	13.285	16.547	19.036

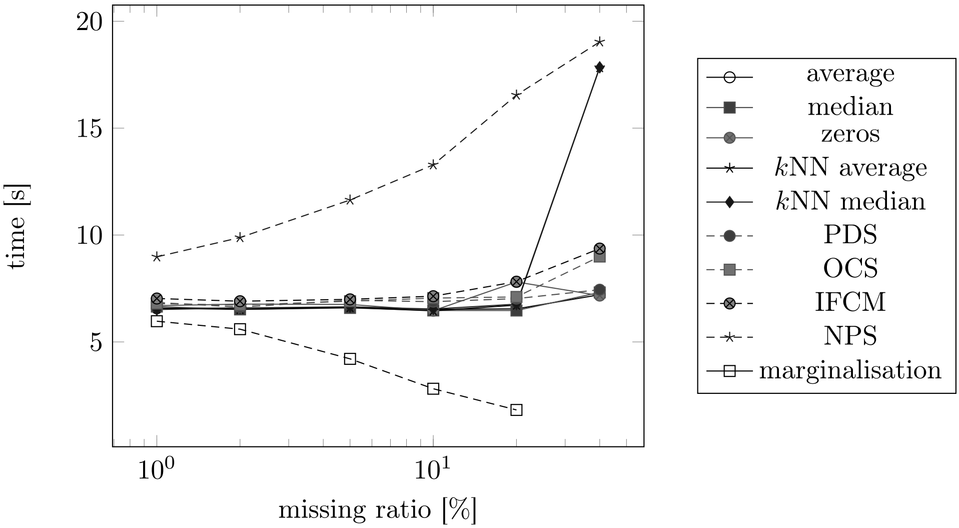

Figure 4. Time (in [s]) of calculation for the 'Gas furnace' data set ($k = 5$).

Table 5

Time (in [s]) of calculations for the 'Concentration of leukocytes' data set, $k = 5$

method	0%	1%	2%	5%	10%	20%	40%
marginalization	9.715	8.506	8.024	6.191	4.608	1.922	–
average	9.141	9.241	9.288	8.787	8.899	8.550	8.589
median	9.650	9.283	9.158	8.729	8.937	8.405	8.645
zeros	9.705	9.472	9.253	9.378	9.733	9.586	8.279
kNN average	9.361	9.248	9.317	8.957	8.964	8.806	9.310
kNN median	9.258	9.251	9.233	8.899	9.400	8.987	9.492
PDS	9.299	9.540	9.612	9.237	9.555	9.347	9.631
OCS	9.569	9.519	9.462	9.304	9.426	9.982	9.621
IFCM	9.290	9.335	9.152	9.585	9.154	9.210	10.439
NPS	11.239	15.221	16.105	43.401	33.267	35.051	33.964

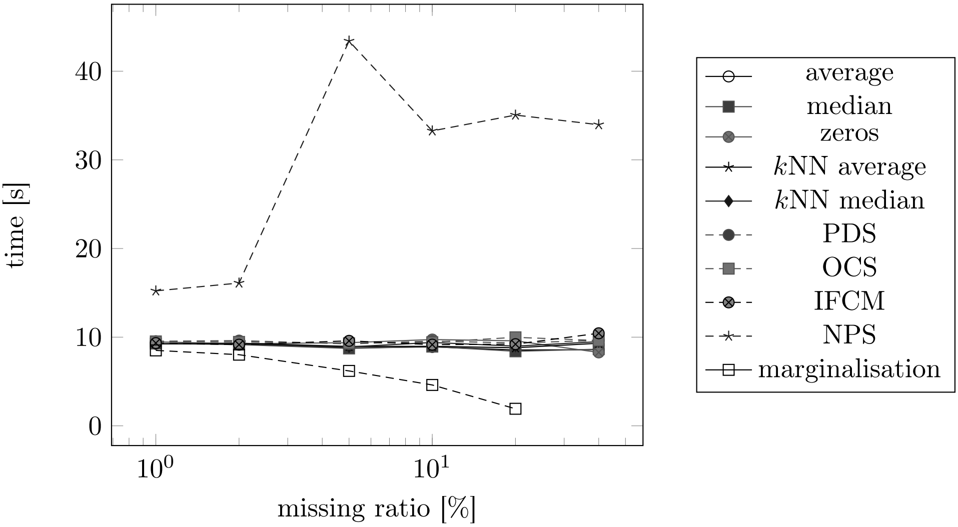

Figure 5. Time (in [s]) of calculations for the 'Concentration of leukocytes' data set, $k = 5$

Table 6
Influence of the k parameter in kNN imputation methods on RSME for the 'Carbon dioxide concentration' data set.

missing	$k = 3$		$k = 5$	
ratio	kNN average	kNN median	kNN average	kNN median
1%	0.32767	0.33307	0.33102	0.32554
2%	0.33595	0.33724	0.32348	0.34835
5%	0.32029	0.36297	0.33797	0.32644
10%	0.32301	0.31562	0.33126	0.33319
20%	0.33810	0.32829	0.39642	0.36547
40%	0.50571	0.52623	0.71197	0.77085

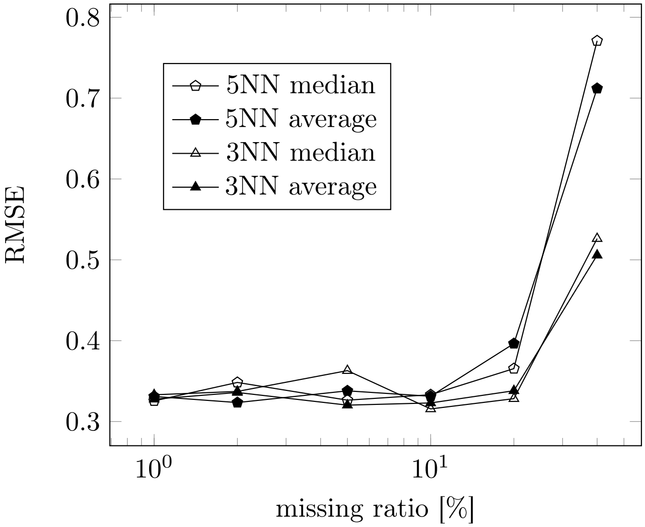

Figure 6. Influence of the k parameter in kNN imputation methods on RSME for the 'Carbon dioxide concentration' data set.

The Imputation with means and medians leads to models with similar generalization ability as the specialized clustering algorithms.

The results of experiments allow for the ordering of techniques from those with the lowest RMSE: kNN median, kNN average, specialized clustering algorithms, median, or average imputation. The marginalization and imputation with a constant value are excluded from the above list, as they are unstable and their efficacy depends on the data set to which they are applied.

An interesting observation from our experiments is the fact that the system is able to create the model with incomplete data that has higher generalization ability (and achieves lower errors) than the model created with complete data. This phenomenon can be observed for the 'Carbon dioxide concentration' and 'Gas furnace'data sets and kNN average and kNN median imputation. It can be also observed for imputing with zeros in the 'Gas furnace' data set. Maybe in these situations, the missing values might be outliers; their values substituted by some mean values.

5. Summary

Neuro-fuzzy systems use three main techniques to create the fuzzy rule base: grid partition, scatter partition (clustering), and hierarchical partition. Most-used is the clustering, although hierarchical partition has some advantages in comparison to clustering.

This paper analyzes the application of the ANNBFIS neuro-fuzzy system with the scatter partition of input domain to incomplete data. We analyzed preprocessing (marginalization of incomplete data, imputation) and specialized clustering algorithms. The created systems were tested with complete data set.

Marginalization is the simplest technique. The creation of fuzzy model with marginalization is the quickest of all analysis approaches. But it is unstable: for some data, it is inappropriate, and for other data, it can be very effective. Imputation of missing values with constant value is similarly unstable.

The best results were obtained with the kNN median and kNN average imputation. The disadvantage of this approach is the long computation time.

The specialized clustering algorithms do not achieve low values of errors. The fuzzy models created with the specialized clustering algorithms do not achieve better values than other techniques. It is worth mentioning that the Nearest Prototype Strategy (NPS) is significantly time consuming in comparison with other approaches.

The imputation with means and medians leads to models with the similar generalization ability as the specialized clustering algorithms.

The interesting observation is the fact that the system is able to create the model with incomplete data that has higher generalization ability (and elaborates lower errors) than the model created with complete data.

Acknowledgements

The authors are grateful to the anonymous referees for their constructive comments that have helped to improve the paper.

References

[1] Acuña E., Rodriguez C.: *The treatment of missing values and its effect in the classifier accuracy.* In: D. Banks, L. House, F. McMorris, P. Arabie, W. G. (eds.), Classification, Clustering and Data Mining Applications, Springer, Berlin, Heidelberg, pp. 639–648. 2004.

[2] Bensaid A. M., Hall L. O., Bezdek J. C., Clarke L. P., Silbiger M. L., Arrington J. A., Murtagh R. F.: *Validity-guided (re)clustering with applications to image segmentation.* In: Transactions on Fuzzy Systems, vol. 4(2), pp. 112–123, 1996. ISSN 1063-6706.

[3] Box G. E. P., Jenkins G.: *Time Series Analysis, Forecasting and Control.* Holden-Day, Incorporated, Oakland, California, 1970.

[4] Cooke M., Green P., Josifovski L., Vizinho A.: *Robust automatic speech recognition with missing and unreliable acoustic data. Speech Communication*, vol. 34, pp. 267–285, 2001.
URL `http://dx.doi.org/10.1016/S0167-6393(00)00034-0`.

[5] Czekalski P.: Evolution-Fuzzy Rule Based System with parameterized consequences. *International Journal of Applied Mathematics and Computer Science*, vol. 16(3), pp. 373–385, 2006.

[6] Czogała E., Łęski J.: *Fuzzy and Neuro-Fuzzy Intelligent Systems.* Series in Fuzziness and Soft Computing. Physica-Verlag, Springer-Verlag Company, Heidelberg, New York, 2000.

[7] Dunn J.C.: A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact, Well Separated Clusters. *Journal Cybernetics*, vol. 3(3), pp. 32–57, 1973.

[8] Ghahramani Z., Jordan M.: *Learning From Incomplete Data.* Tech. rep., Lab Memo No. 1509, CBCL Paper No. 108, MIT AI Lab, 1995.

[9] Grzymała-Busse J., Goodwin L., Grzymala-Busse W., Zheng X.: *Handling Missing Attribute Values in Preterm Birth Data Sets.* D. Slezak, J. Yao, J. Peters, W. Ziarko, X. Hu, (eds.), Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, Lecture Notes in Computer Science, vol. 3642, pp. 342–351. Springer Berlin / Heidelberg, 2005. ISBN 978-3-540-28660-8.

[10] Grzymała-Busse J., Hu M.: *A Comparison of Several Approaches to Missing Attribute Values in Data Mining.* In: W. Ziarko, Y. Yao, (eds.), Rough Sets and Current Trends in Computing, Lecture Notes in Computer Science, vol. 2005, pp. 378–385. Springer Berlin / Heidelberg, 2001. ISBN 978-3-540-43074-2.

[11] Hathaway R., Bezdek J.: Fuzzy c-means clustering of incomplete data. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 31(5), pp. 735–744, 2001. ISSN 1083-4419.
URL `http://dx.doi.org/10.1109/3477.956035`.

[12] Jang J. S. R.: ANFIS: Adaptive-Network-Based Fuzzy Inference System. *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 23(3), pp. 665–684, 1993.

[13] Kalton G., Kasprzyk D.: The treatment of missing survey data. *Survey Methodology*, vol. 12, pp. 1–16, 1986.

[14] Łęski J.: *Systemy neuronowo-rozmyte (Neuro-fuzzy systems)*. Wydawnictwa Naukowo-Techniczne, Warszawa, 2008. ISBN 978-83-204-3229-9.

[15] Łęski J., Czogała E.: A new artificial neural network based fuzzy inference system with moving consequents in if-then rules and selected applications. *Fuzzy Sets and Systems*, vol. 108(3), pp. 289–297, 1999. ISSN 0165-0114.
URL `http://dx.doi.org/10.1016/S0165-0114(97)00314-X`.

[16] Mackey M. C., Glass L.: Oscillation and chaos in physiological control systems. *Science*, vol. 197(4300), pp. 287–289, 1977.

[17] Matyja A., Simiński K.: Comparison of algorithms for clustering incomplete data. *Foundations of Computing and Decision Sciences*, vol. 39(2), pp. 107–127, 2014.
URL `http://dx.doi.org/10.2478/fcds-2014-0007`.

[18] Mundfrom D.J., Whitcomb A.: Imputing Missing Values: The Effect on the Accuracy of Classification. *Multiple Linear Regression Viewpoints*, vol. 25(1), pp. 13–19, 1998.

[19] Nelles O., Fink A., Babuška R., Setnes M.: Comparison of Two Construction Algorithms for Takagi-Sugeno Fuzzy Models. *International Journal of Applied Mathematics and Computer Science*, vol. 10(4), pp. 835–855, 2000.

[20] Nelles O., Isermann R.: Basis function networks for interpolation of local linear models. *Proceedings of the 35th IEEE Conference on Decision and Control*, vol. 1, pp. 470–475, 1996.

[21] Pal N. R., Bezdek J. C.: On cluster validity for the fuzzy c-means model. *Fuzzy Systems, IEEE Transactions on*, vol. 3(3), pp. 370–379, 1995.

[22] Reichenbach H.: Wahrscheinlichkeitslogik. *Erkenntnis*, vol. 5, pp. 37–43, 1935. ISSN 0165-0106. URL `http://dx.doi.org/10.1007/BF00172280`.

[23] Rubin D.: *Multiple Imputation For Nonresponse In Surveys*. John Wiley & Sons, Inc., 1987.

[24] Sikora M., Krzystanek Z., Bojko B., Śpiechowicz K.: Application of a hybrid method of machine learning for description and on-line estimation of methane hazard in mine workings. *Journal of Mining Sciences*, vol. 47(4), pp. 493–505, 2011.

[25] Simiński K.: Neuro-fuzzy system with hierarchical domain partition. In: *Proceedings of the International Conference on Computational Intelligence for Modelling, Control and Automation (CIMCA 2008)*, pp. 392–397. IEEE Computer

Society Publishing, Vienna, Austria, 2008. ISBN 978-0-7695-3514-2. URL `http://dx.doi.org/10.1109/CIMCA.2008.67`.

[26] Simiński K.: *Patchwork neuro-fuzzy system with hierarchical domain partition.* In: M. Kurzyński, M. Woźniak (eds.), Computer Recognition Systems 3, Advances in Intelligent and Soft Computing, vol. 57, pp. 11–18. Springer-Verlag, Berlin, Heidelberg, 2009. URL `http://dx.doi.org/10.1007/978-3-540-93905-4_2`.

[27] Simiński K.: Neuro-rough-fuzzy approach for regression modelling from missing data. *International Journal of Applied Mathematics and Computer Science*, vol. 22(2), pp. 461–476, 2012. URL `http://dx.doi.org/DOI:10.2478/v10006-012-0035-4`.

[28] Simiński K.: Clustering with missing values. *Fundamenta Informaticae*, vol. 123(3), pp. 331–350, 2013.

[29] Simiński K.: Rough fuzzy subspace clustering for data with missing values. *Computing & Informatics*, vol. 33(1), pp. 131–153, 2014.

[30] Simiński K.: Rough subspace neuro-fuzzy system. *Fuzzy Sets and Systems*, 2014. ISSN 0165-0114. URL `http://dx.doi.org/http://dx.doi.org/10.1016/j.fss.2014.07.003`.

[31] Timm H., Döring C., Kruse R.: Different approaches to fuzzy clustering of incomplete datasets. *International Journal of Approximate Reasoning*, vol. 35(3), pp. 239–249, 2004. ISSN 0888-613X. URL `http://dx.doi.org/DOI:10.1016/j.ijar.2003.08.004`. Integration of Methods and Hybrid Systems.

[32] Timm H., Kruse R.: Fuzzy cluster analysis with missing values. *NAFIPS 1998 Conference of the North American Fuzzy Information Processing Society*, pp. 242–246. 1998. URL `http://dx.doi.org/10.1109/NAFIPS.1998.715573`.

[33] Troyanskaya O., Cantor M., Sherlock G., Brown P., Hastie T., Tibshirani R., Botstein D., Altman R.B.: Missing value estimation methods for DNA microarrays. *Bioinformatics*, vol. 17(6), pp. 520–525, 2001. URL `http://dx.doi.org/10.1093/bioinformatics/17.6.520`.

[34] Wagstaff K. L., Laidler V. G.: Making the Most of Missing Values: Object Clustering with Partial Data in Astronomy. *Proceedings of Astronomical Data Analysis Software and Systems XIV*, vol. 347, pp. 172–176. Pasadena, California, USA, 2005.

[35] Xie X., Beni G.: A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13(8), pp. 841–847, 1991.

[36] Zhang C., Zhu X., Zhang J., Qin Y., Zhang S.: GBKII: An Imputation Method for Missing Values. *Advances in Knowledge Discovery and Data Mining*, vol. 4426, pp. 1080–1087, 2007.

[37] Zhang S.: Shell-neighbor method and its application in missing data imputation. In: *Applied Intelligence*, vol. 35(1), pp. 123–133, 2011. ISSN 0924-669X. URL `http://dx.doi.org/10.1007/s10489-009-0207-6`.

Affiliations

Marcin Sikora
Independent researcher

Krzysztof Simiński
Silesian University of Technology, Faculty of Automatic Control, Electronics and Computer Science, Krzysztof.Siminski@polsl.pl

Received: 24.06.2014
Revised: 8.08.2014
Accepted: 26.08.2014