# Group Lasso based redundancy-controlled feature selection for fuzzy neural network*

**YANG Jun, XU Yongyong, WANG Bin, LI Bo, HUANG Ming, and GAO Tao****
*The 15th Research Institute of China Electronics Technology Group Corporation, Beijing 100191, China*

(Received 1 April 2022; Revised 11 January 2023)


If there are a lot of inputs, the readability of the "If-then" fuzzy rule is reduced, and the complexity of the fuzzy neural network structure will be increased. Hence, to optimize the structure of the fuzzy rule based neural network, a group Lasso based redundancy-controlled feature selection (input pruning) method is proposed. For realizing feature selection, the linear/nonlinear redundancy between features is considered, and the Pearson's correlation coefficient is employed to construct the additive redundancy-controlled regularizer in the error function. In addition, considering the past gradient information, a novel parameter optimization method is presented. Finally, we demonstrate the effectiveness of our method on two benchmark classification datasets.
**Document code:** A **Article ID:** 1673-1905(2023)05-0284-6
**DOI** https://doi.org/10.1007/s11801-023-2053-x

To understand the world better and make decisions, human beings always eager to collect more and more data at any time, especially in the digital/big-data era. It is inevitable that the data is increasing drastically both row-wise and column-wise, which directly leads to the curse of dimensionality. In different areas such as healthcare, bioinformatics, transportation, social media and online education, the so-called scenario curse of dimensionality always exists[1].

One method to solve the curse of dimensionality is feature selection. Feature selection is a process of selecting a feature subset from the original feature set based on feature importance and feature correlation. Features can be classified by different methods, and redundant features are considered as useful features, but they are with linear/nonlinear relation between each other[2].

Feature selection method mainly includes filter methods, wrapper methods, and integrated/embedded methods. The filter methods ignore the performance feedback of the modeling algorithms which will use the selected features. The wrapper methods evaluate the importance of the selected features using the modeling algorithms that will finally employ the selected features. It is obvious that the wrapper methods may yield better application effect, for different modeling algorithms are not likely to correspond to one best set of selected features. But when faced with high dimensional problem, it will be time-consuming to assess all possible subsets of features. The integrated/embedded feature selection methods integrate the feature selection process and the training process into one single unified optimization procedure. In addition, they are time-saving, for the modeling algorithms do not need to evaluate all the possible subsets.

As one technique of computational intelligence, fuzzy systems can be divided into type-1 and type-2 systems, and type-1/type-2 fuzzy systems include Mamdani-Assilian (MA) and Takagi-Sugeno (TS) models[3]. Fuzzy rule based models are widely used to realize feature selection. POMARES et al[4] proposed a type-1 fuzzy model based method that can select the input variables and determine the number of membership functions. Combined with fuzzy model, a feature-weighted detector network is presented[5], and this strategy evaluates the importance of features using the values of connection weights. Exponential function based feature selection method is introduced[6], which is applied to type-1 fuzzy models. Inspired by the exponential function, recurrent feature selection fuzzy neural network is designed to realize intelligent back stepping control[7]. Using soft version minimum T-norm, the type-1 fuzzy neural network is firstly used to realize feature selection for high dimensional classification problems[8]. Compared with type-1 fuzzy models, type-2 fuzzy models are beneficial to deal with uncertainty. Interval type-2 fuzzy neural network is used to realize feature selection for noised data[9]. For the interval type-2 fuzzy system, a different exponential function is used to adjust the membership function to realize feature selection[10]. The above mentioned feature selection methods usually introduce additional parameters for the original fuzzy system, which increases the model's complexity. To overcome this problem, in our previous study we use the internal parameter to realize feature selection[11].

* This work has been supported by the Major Project of National Natural Science Foundation of China (No.U19B2019).
** E-mail: gaotao_1989@126.com

Among the selected features, it is apparent that different features may depend on each other. But there is limited research about feature dependence in the feature selection process for fuzzy models. A fuzzy rule-based framework is firstly used to design the method to control the feature redundancy[12]. A fast correlation-based filter is introduced to consider the feature relevance and redundancy using entropy[13]. The minimum redundancy-maximum relevance method uses a filter based technique to realize feature selection, where the feature has the maximum relevance and minimum redundancy[14]. Be similar to the minimum redundancy-maximum relevance strategy, other mutual information based redundancy-controlled feature selection methods are presented[15,16]. To address the problem of redundancy-controlled feature selection, some methods using trace-based class separability[17], support vector machine (SVM)[18], fuzzy entropy[19] and Pearson correlation coefficient[20,21] are introduced.

To design the feature selection method for the fuzzy neural network (FS-FNN), in our previous work[11], we propose an internal parameter based gate function, which does not introduce the additional parameter as the exponential function based method[6]. Using the gate-function curve[11], we analyze the feasibility of feature selection strategy. In addition, the reason for designing group Lasso based regularizer is presented, i.e., one feature corresponds to different numbers of gate functions, and compressing one gate width value cannot guarantee the realization of feature selection, for other nonzero widths of the bad feature influence the value of firing strength, which contradicts with the core feature selection idea. Actually, in the selected feature sets using group Lasso, there may exist redundant features among the good features. Redundant features certainly hinder the structure simplification of FNN, but proper redundant features can improve the robustness of the network, so it is necessary to control the level of redundant features. Hence, using the Pearson's correlation coefficient, the group Lasso based redundancy-controlled feature selection method for fuzzy neural network (RCFS-FNN) is proposed in this paper, and the goal of our feature selection strategy is selecting the essential features with a number of redundant features. In addition, using the past gradient information, we design a new learning algorithm to realize parameter updating.

For the $q$-classes problem, supposing the original dataset is $\{\boldsymbol{x}_j^*, s_j\}_{j=1}^{J} \subset \mathbb{R}^t \times \mathbb{R}$, where $\boldsymbol{x}_j^*$ denotes the attribute vector of the $j$th sample, $s_j$ is the corresponding class label and $J$ corresponds to the total number of all the samples. In general, $\boldsymbol{x}_j^*$ and $s_j$ are firstly transformed using some normalization method and label quantification method. In this paper, we separately employ the maximum-minimum value method and one-hot coding method to get the preprocessed dataset $\{\boldsymbol{x}_j, \boldsymbol{o}_j\} \subset \mathbb{R}^t \times \mathbb{R}^q$, where $\boldsymbol{x} = (x_1, x_2, ..., x_t)^{\mathrm{T}} \in \mathbb{R}^t$, $\boldsymbol{o}=(o_1,o_2,...,o_q)^{\mathrm{T}} \in \mathbb{R}^q$, $x_p = (x_p^* - x_{p,\min}^*)/(x_{p,\max}^* - x_{p,\min}^*) \in [0,1]$, and $o_s$=1, $o_{l\neq s}$=0. To get the following If-then rule, we use fuzzy c-means (FCM) clustering algorithm to cluster all the randomly selected training augmented vectors $(\boldsymbol{x}_j^{\mathrm{T}}, \boldsymbol{o}_j^{\mathrm{T}})^{\mathrm{T}}$, where $j$=1, 2,…, $J_{\mathrm{Tr}}$, and the detailed procedure is presented in our previous work[11]. The core idea of clustering algorithm based rule initialization method is that one clustering center is translated into one fuzzy rule.

$$R_{i_s}: \text{If } x_1 \text{ is } A_{i1} \text{ and } x_2 \text{ is } A_{i2} \text{ ... and } x_t \text{ is } A_{it}, \text{ then } y_s \text{ is } u_{is}. \quad (1)$$

Fig.1 presents the FNN based multiple inputs and multiple outputs classifier which is a four-layers model, and the $i$th rule denoted using layer nodes in Fig.1 realizes one local mapping relation between $\boldsymbol{x}$ and $\boldsymbol{o}$. The introduction of each layer is introduced as below.

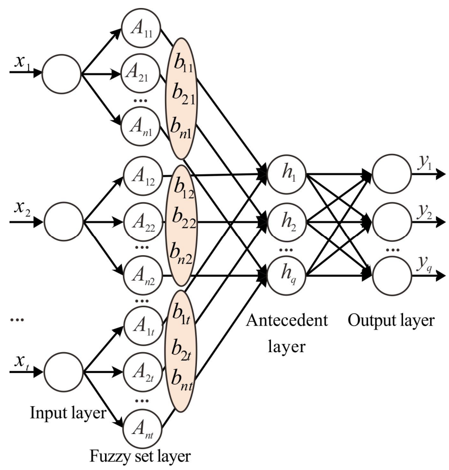


**Fig.1 Structure of the FNN based classifier**

**Input layer**: For the input linguistic attribute vector $\boldsymbol{x} = (x_1, x_2, ..., x_t)^{\mathrm{T}} \in \mathbb{R}^t$, there are $t$ nodes, and the $p$th node maps the linguistic variable $x_p$ (usually called as input feature) directly.

**Fuzzy set layer**: Fuzzy set $A_{ip} = \{(x_p, \mu_{A_{ip}}(x_p)) | x_p \in U\}$ is used as the linguistic value to describe the linguistic variable $x_p$. For the clustering algorithm based rule initialization method, the fuzzy sets' numbers of different linguistic variables are identical, which are same as the number of clustering centers $n$, i.e., for $x_p$, there are $n$ fuzzy sets. For each fuzzy set $A_{ip}$, the value of $\mu_{A_{ip}}(x_p)$ reflects the degree of $x_p$ belonging to the fuzzy sets $A_{ip}$. $\mu_{A_{ip}}(x_p) \in [0,1]$ is called membership function and modeled by the following expression:

$$\mu_{A_{ip}}(x_p, a_{ip}, b_{ip}) = \exp(-(x_p - a_{ip})^2 b_{ip}^2). \quad (2)$$

Actually, Eq.(2) is a variation of Gaussian membership function, where $b_{ip}$=1/$\sigma_{ip}$ is the adaptive parameter to realize feature selection, and $a_{ip}$ denotes the Gaussian center.

**Antecedent layer**: Each node in the antecedent layer realizes intersection operation for all the antecedent parts of one rule using Eq.(3) or Eq.(4)

$$h_i = \min\{\mu_{A_{i1}}, \mu_{A_{i2}}, ..., \mu_{A_{it}}\}, \tag{3}$$

$$h_i = \prod_{p=1}^{t} \mu_{A_{ip}}. \tag{4}$$

Both T-norm expressions can be used to design the feature selection method, for the membership value 1 of bad features does not influence the firing strength value $h_i$, i.e., $\min\{\mu_{ip}, 1\}=\mu_{ip}$ and $\mu_{ip}\cdot 1=\mu_{ip}$. In this paper, we use the gradient based technique to realize feature selection, which requires the differentiability of expressions. Eq.(3) is not differential, so Eq.(4) is borrowed to compute the value of $h_i$.

**Output layer**: To solve the real problems, the fuzzy firing strength value $h_i$ should be defuzzified. The traditional defuzzification methods are maximum method and gravity method as follows respectively.

$$y_s = \max\{h_1, h_2, \cdots, h_n\}, \tag{5}$$

$$y_s = \sum_{i=1}^{n} \left(h_i \Big/ \sum_{i=1}^{n} h_i\right) v_{is}. \tag{6}$$

The maximum method also faces the problem of non-differentiability, so we employ the gravity method and simplify it as below:

$$y_s = \sum_{i=1}^{n} h_i v_{is}. \tag{7}$$

For feature selection, we employ our previous strategy as below[11],

$$E(\boldsymbol{w}) + \lambda \sum_{p=1}^{t} \|\boldsymbol{b}_p\| = \frac{1}{2} \sum_{j=1}^{J_{Tr}} \sum_{s=1}^{q} (y_{js} - o_{js})^2 + \lambda \sum_{p=1}^{t} \|\boldsymbol{b}_p\|, \tag{8}$$

where $\boldsymbol{w}$ denotes the argument of loss function consisting of antecedent and consequent parameters in fuzzy rule base, i.e., $\boldsymbol{w}=(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_t, \boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_t, \boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_q)$, $\boldsymbol{a}_p=(a_{1p}, a_{2p}, \ldots, a_{np})$, $\boldsymbol{b}_p=(b_{1p}, b_{2p}, \ldots, b_{np})$, $\boldsymbol{v}_s=(v_{1s}, v_{2s}, \ldots, v_{ns})$. Neural network based learning algorithm realizes the rule parameter optimization. The regularization term coefficient $\lambda$ balances the classification performance and the feature selection performance in the learning procedure.

For evaluating the redundancy of different features, the Pearson's relevance coefficient is used to calculate the redundancy, and the relationship between the features and class labels can also be calculated using

$$\rho(x_p, x_l) = \mathrm{Cov}(x_p, x_l) \Big/ \left(\sqrt{\mathrm{Var}(x_p)}\sqrt{\mathrm{Var}(x_l)}\right) =$$

$$\sum_{j=1}^{J_{Tr}} (x_{jp} - \bar{x}_p)(x_{jl} - \bar{x}_l) \Big/ \left(\sqrt{\sum_{j=1}^{J_{Tr}} (x_{jp} - \bar{x}_p)^2} \sqrt{\sum_{j=1}^{J_{Tr}} (x_{jl} - \bar{x}_l)^2}\right). \tag{9}$$

Based on the feature selection strategy, the values of $\|\boldsymbol{b}_p\|$s ($p$=1, 2, …, $t$) for good features are big. To control the redundancy, we should simultaneously control the values of $\|\boldsymbol{b}_l\|$s ($l$=1, 2, …, $t$, $l\neq p$) corresponding to the redundant features in the feature selection procedure. Hence, for $\|\boldsymbol{b}_p\|$, another additive regularization term containing $(\|\boldsymbol{b}_p\| \sum_{l=1, l\neq p}^{t} \|\boldsymbol{b}_l\|)/(t-1)$ is designed as below to control the redundancy using the penalty coefficient $\beta$. In the meanwhile, the Pearson's relevance coefficient based multiplied parameter $\rho^2(x_p, x_l)$ is included in the additive regularizer. Considering all the features, to realize feature selection, we employ our previous strategy as below[22],

$$E^*(\boldsymbol{w}) = E(\boldsymbol{w}) + \lambda \sum_{p=1}^{t} \|\boldsymbol{b}_p\| +$$

$$\beta \frac{1}{t(t-1)} \sum_{p=1}^{t} \|\boldsymbol{b}_p\| \sum_{l\neq p, l=1}^{t} \|\boldsymbol{b}_l\| \rho^2(x_p, x_l). \tag{10}$$

Considering the influence of past information on the current state, we propose a gradient memory based learning algorithm as below,

$$\boldsymbol{w}^{k+1} = \boldsymbol{w}^k + \eta \begin{cases} -E_{\boldsymbol{w}}^*(\boldsymbol{w}^k), k=0; \\ -E_{\boldsymbol{w}}^*(\boldsymbol{w}^k) - \gamma E_{\boldsymbol{w}}^*(\boldsymbol{w}^{k-1}), k\geq 1, \end{cases} \tag{11}$$

where $\boldsymbol{w}^0$ is determined by the clustering algorithm, $\eta$ is the learning rate, $E_{\boldsymbol{w}}^*(\boldsymbol{w}^k)$ denotes the gradient, and $\gamma$ is used to realize the gradient memory. Differing from the method in Ref.[22], we update $\boldsymbol{b}_p$ and $\boldsymbol{v}_s$ simultaneously.

To evaluate the performance of the proposed redundancy-controlled feature selection method, the classification accuracy as below is used as the evaluation criterion.

$$Accuracy = \sum_{j=1}^{J} L_j / J \cdot 100\%, \tag{12}$$

where $L_j$=1 if the $j$th sample is classified into the correct class, otherwise $L_j$=0.

Two benchmark classification problems, Iris and Wisconsin Breast Cancer (WBC), are borrowed to test the effectiveness of the proposed feature selection strategy. Tab.1 summarizes the detail of the two datasets, including the number of input features, number of samples and classes. The two datasets can be downloaded from the University of California at Irvine (UCI)[23]. For each dataset, 50 trials are conducted. In each trial, 60% samples are randomly selected as the training samples, and the rest for the test. All the simulations are carried out in MATLAB R2018b Intel(R) Core(TM) i7-9700 CPU.

**Tab.1 Summary of the datasets used**

| Datasets | Number of input features | Number of samples | Classes |
|---|---|---|---|
| Iris | 4 | 150 (50+50+50) | 3 |
| WBC | 9 | 683 (444+239) | 2 |

In this part, the simulation results of Iris are reported. Using Eq.(9), the correlation matrix for Iris is presented in Tab.2. Observing this table, we find that feature 3 and feature 4 are with high correlation with the classes, and the Pearson's relevance coefficient $\rho(x_3, x_4)$ is 0.97, which means that features 3 and 4 are important and correlated features.

To test the validation of RCFS-FNN, when $\eta$=0.005 and $\lambda$=1.5, the simulation results of FS-FNN and RCFS-FNN are reported in Tab.3. It is easy to see that the redundancy-controlled feature selection method avoids selecting the high correlated features in each trial, for $s_{3,4}$=0, which means that with a proper penalty coefficient $\beta$, our proposed feature selection method can control the number of redundant features. In addition, using a different $\lambda$, another 50 trials are conducted, and the similar feature selection results are obtained, i.e., the RCFS-FNN can control the redundancy in the process of feature selection.

**Tab.2 Correlation matrix for Iris dataset, where *FLCor* denotes the correlation between features and class labels**

| *F* | 1 | 2 | 3 | 4 | *FLCor* |
|---|---|---|---|---|---|
| 1 | 1.00 | −0.11 | 0.88 | 0.82 | 0.79 |
| 2 | −0.11 | 1.00 | −0.42 | −0.36 | −0.42 |
| 3 | 0.88 | −0.42 | 1.00 | **0.97** | **0.96** |
| 4 | 0.82 | −0.36 | **0.97** | 1.00 | **0.96** |

**Tab.3 Comparison of feature selection results between FS-FNN and RCFS-FNN for the Iris dataset**

| Methods | FS-FNN | RCFS-FNN | FS-FNN | RCFS-FNN |
|---|---|---|---|---|
| $\eta$ | 0.005 | 0.005 | 0.005 | 0.005 |
| $\lambda$ | 1.5 | 1.5 | 2 | 2 |
| $\beta$ | 0 | 80 | 0 | 80 |
| $s_3$ | 4 | 13 | 7 | 15 |
| $s_4$ | 25 | 37 | 27 | 35 |
| $s_{3,4}$ | 21 | 0 | 16 | 0 |
| $s_{\text{others}}$ | 0 | 0 | 0 | 0 |
| *F*1 | 0 | 0 | 0 | 0 |
| *F*2 | 0 | 0 | 0 | 0 |
| *F*3 | 25 | 13 | 23 | 15 |
| *F*4 | 46 | 37 | 43 | 35 |
| AveSel | 1.42 | 1 | 1.32 | 1 |
| AveTrAcc | 94.69 | 95.53 | 96.02 | 96.24 |
| AveTeAcc | 94.47 | 95.90 | 94.63 | 94.87 |

In Tab.4, the results about the effect of redundancy-controlled factor $\beta$ on the correlation of different features are recorded. From this table, we find that the average maximum correlation of selected features is decreased with the increasing of $\beta$. The curves for Iris plotted in Fig.2 illustrate that RCFS-FNN makes the norms of bad feature and redundant features tend to zero in the optimization process.

**Tab.4 Simulation results of different *β* for Iris**

| Datasets ($\eta$, $\lambda$) | $\beta$ | Ave SelFea | Ave MaxCor | Ave TrAcc | Ave TeAcc | $n$ |
|---|---|---|---|---|---|---|
| Iris | 0 | 2.78 | 0.92 | 0.97 | 0.96 | |
| (0.005,1.2) | 50 | 1.74 | 0.63 | 0.97 | 0.96 | 3 |
| | 200 | 1.10 | 0.04 | 0.96 | 0.96 | |

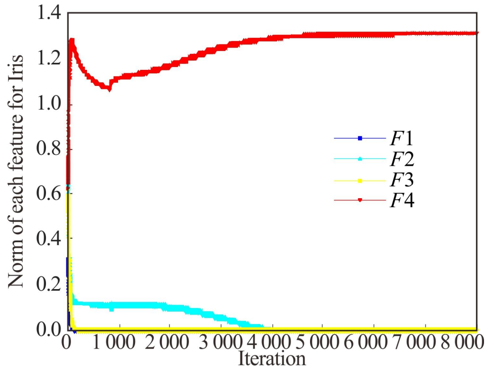


**Fig.2 Norm variation of each feature for Iris ($\eta$=0.005, $\lambda$=1.2, $\beta$=200, $n$=3)**

For WBC, in Tab.5, the numbers of selected features with different $\beta$ are summarized. The number of feature 5 and feature 9 is 0, respectively, which leads us to conclude that the two features are not important. In addition, using Tab.5, we find that the high correlated features, feature 2 and feature 3, are not selected simultaneously when $\beta$=1 000, which proves the effectiveness of RCFS-FNN.

**Tab.5 Effects of *β* on the number of selected features for WBC when *η*=0.008 and *λ*=0.55**

| $\beta$ | *F*1 | *F*2 | *F*3 | *F*4 | *F*5 | *F*6 | *F*7 | *F*8 | *F*9 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 26 | 40 | 15 | 1 | 0 | 50 | 4 | 2 | 0 |
| 500 | 8 | 6 | 6 | 0 | 0 | 49 | 0 | 1 | 0 |
| 1 000 | 0 | 0 | 2 | 0 | 0 | 44 | 0 | 1 | 0 |

Tab.6 shows the average performance of the proposed model for WBC dataset. Observing it, we see that with the increasing of redundancy-controlled factor $\beta$, the number of redundant features will be reduced, and the average maximum correlation is minimized. From Fig.3, it is easy to see that RCFS-FNN selects feature 6 in the feature selection procedure, for the norm of F6 converges to a nonzero value.

**Tab.6 Average performance of RCFS-FNN for WBC dataset**

| Datasets ($\eta$, $\lambda$) | $\beta$ | Ave SelFea | Ave MaxCor | Ave TrAcc | Ave TeAcc | $n$ |
|---|---|---|---|---|---|---|
| WBC | 0 | 4.54 | 0.82 | 96.4 | 95.5 | |
| (0.008,0.05) | 150 | 2.24 | 0.64 | 95.8 | 95.1 | 3 |
| | 800 | 1.22 | 0.12 | 91.6 | 91.6 | |

Comparisons with RCFS, mFSMLP-CoR, SGLC[21] and our proposed method on Iris and WBC are given in Tab.7. From this table, it is easy to see that the average maximum correlation of RCFS-FNN is lower than other models, which indicates that the proposed model can effectively control the redundancy in the selected feature subset. In addition, the classification performance of RCFS-FNN is better than its competitors.

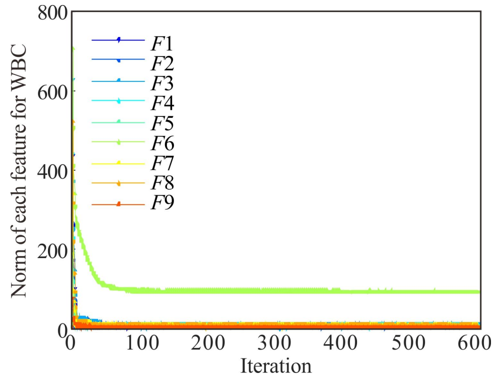


**Fig.3 Norm variation of each feature for WBC ($\eta$=0.008, $\lambda$=0.05, $\beta$=800, $n$=3)**

**Tab.7 Comparison with other models**

| Models | AccTe | | MaxCor | | AveMaxCor | |
|---|---|---|---|---|---|---|
| | Iris | WBC | Iris | WBC | Iris | WBC |
| RCFS | 95.4 | 94.6 | 0.96 | 0.90 | * | * |
| mFSMLP-CoR | 94.8 | 95.3 | 0.42 | 0.69 | * | * |
| SGLC | 96.0 | 96.0 | 0.96 | 0.69 | 0.42 | 0.64 |
| RCFS-FNN | **96.4** | **96.2** | 0.96 | 0.76 | **0.41** | **0.60** |

To control the redundancy of the selected subsets in feature selection process, the Pearson's correlation coefficient based feature selection strategy is proposed. In addition, using a gradient memory based learning algorithm, the parameters of antecedent and consequent part are updated. Experiments demonstrate the validation of the redundancy-controlled feature selection model. It is a pity that the effectiveness of our method for high dimensional problems with thousands of inputs has not been studied, which pushes us to study it in the future work.

## Statements and Declarations

The authors declare that there are no conflicts of interest related to this article.

## References


[1] BUGATA P, DROTAR P. Feature selection based on a sparse neural-network layer with normalizing constraints[J]. IEEE transactions on cybernetics, 2021.

[2] VENKATESH B, ANURADHA J. A review of feature selection and its methods[J]. Cybernetics and information technologies, 2019, 19(1)：3-26.

[3] WU D R, ZENG Z G, MO H, et al. Interval type-2 fuzzy sets and systems：overview and outlook[J]. Acta automatica sinica, 2020, 46(8)：1539-1556.

[4] POMARES H, ROJAS I, GONZÁLEZ J, et al. Structure identification in complete rule-based fuzzy systems[J]. IEEE transactions on fuzzy systems, 2002, 10(3)：349-359.

[5] LI R P, MUKAIDONO M, TURKSEN I B. A fuzzy neural network for pattern classification and feature selection[J]. Fuzzy sets and systems, 2002, 130(1)：101-108.

[6] CHEN Y C, PAL N R, CHUNG I F. An integrated mechanism for feature selection and fuzzy rule extraction for classification[J]. IEEE transactions on fuzzy systems, 2011, 20(4)：683-698.

[7] LIN F J, CHEN S G, HSU C W. Intelligent backstepping control using recurrent feature selection fuzzy neural network for synchronous reluctance motor position servo drive system[J]. IEEE transactions on fuzzy systems, 2018, 27(3)：413-427.

[8] XUE G, CHANG Q, WANG J, et al. An adaptive neuro-fuzzy system with integrated feature selection and rule extraction for high-dimensional classification problems[EB/OL]. (2022-01-28) [2022-03-30]. arXiv preprint arXiv：2201.03187, 2022.

[9] LIN C T, PAL N R, WU S L, et al. An interval type-2 neural fuzzy system for online system identification and feature elimination[J]. IEEE transactions on neural networks and learning systems, 2014, 26(7)：1442-1455.

[10] HOU S, CHU Y, FEI J. Adaptive type-2 fuzzy neural network inherited terminal sliding mode control for power quality improvement[J]. IEEE transactions on industrial informatics, 2021, 17(11)：7564-7574.

[11] GAO T, BAI X, ZHANG L, et al. Feature selection for fuzzy neural networks using group Lasso regularization[C]//2021 IEEE Symposium Series on Computational Intelligence (SSCI), December 5-7, 2021, Orlando, FL, USA. New York：IEEE, 2021：1-8.

[12] CHUNG I F, CHEN Y C, PAL N R. Feature selection with controlled redundancy in a fuzzy rule based framework[J]. IEEE transactions on fuzzy systems, 2017, 26(2)：734-748.

[13] YU L, LIU H. Efficient feature selection via analysis of relevance and redundancy[J]. The journal of machine learning research, 2004, 5：1205-1224.

[14] PENG H, LONG F, DING C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy[J]. IEEE transactions on pattern analysis and machine intelligence, 2005, 27(8)：1226-1238.

[15] BATTITI R. Using mutual information for selecting features in supervised neural net learning[J]. IEEE transactions on neural networks, 1994, 5(4)：537-550.

[16] ESTÉVEZ P A, TESMER M, PEREZ C A, et al. Normalized mutual information feature selection[J]. IEEE transactions on neural networks, 2009, 20(2)：189-201.

[17] ZHOU L, WANG L, SHEN C. Feature selection with redundancy-constrained class separability[J]. IEEE transactions on neural networks, 2010, 21(5)：853-858.

[18] AKSU Y, MILLER D J, KESIDIS G, et al. Margin-maximizing feature elimination methods for linear and nonlinear kernel-based discriminant functions[J]. IEEE transactions on neural networks, 2010, 21(5)：701-717.

[19] LEE H M, CHEN C M, CHEN J M, et al. An efficient fuzzy classifier with feature selection based on fuzzy entropy[J]. IEEE transactions on systems, man, and

cybernetics, part B (cybernetics), 2001, 31(3)：426-432.
[20] PAL N R, MALPANI M. Redundancy-constrained feature selection with radial basis function networks[C]//2012 International Joint Conference on Neural Networks (IJCNN), June 10-15, 2012, Brisbane, QLD, Australia. New York：IEEE, 2012：1-8.
[21] WANG J, ZHANG H, WANG J, et al. Feature selection using a neural network with group Lasso regularization and controlled redundancy[J]. IEEE transactions on neural networks and learning systems, 2021, 32(3)：1110-1123.
[22] GAO T, BAI X, ZHANG L, et al. Redundancy-controlled feature selection for fuzzy neural networks[C]//2021 International Conference on High Performance Big Data and Intelligent Systems (HPBD&IS), December 5-7, 2021, Shenzhen, China. New York：IEEE, 2021：172-176.
[23] School of Information and Computer Science, University of California. UCI machine learning repository[EB/OL]. [2022-03-30]. http：//archive.ics.uci.edu/ml.