*Tạp chí Khoa học và Công nghệ **53** (5) (2015) 583-597*

AN APPLICATION OF FEATURE SELECTION FOR THE FUZZY RULE BASED CLASSIFIER DESIGN BASED ON AN ENLARGED HEDGE ALGEBRAS FOR HIGH-DIMENSIONAL DATASETS

Pham Dinh Phong[1, 2]

[1]*Prévoir Vietnam, 23 Phan Chu Trinh, Hanoi, Vietnam*

[2]*University of Engineering and Technology, Hanoi National University, 144 Xuan Thuy, Cau Giay, Hanoi, Vietnam*

Email: *dinhphongpham@gmail.com*

Received: 6 January 2015; Accepted for publication: 19 April 2015

ABSTRACT

The fuzzy rule based classification system (FRBCS) design methods, whose fuzzy rules are in the form of if-then sentences, have been being studied intensively during last years. One of the eminent FRBCS design methods utilizing an enlarged hedge algebras as a formal mechanism to design optimal linguistic terms integrated with their trapezoidal fuzzy sets has been proposed by Ho N. C. et. al. As the other methods, an entanglement of this approach needed to be solved is dealing with the high-dimensional and multi-instance datasets. This paper presents an approach to tackle the high-dimensional dataset problem for the FRBCS design method based on an enlarged hedge algebras by utilizing the feature selection algorithm proposed by Sun X. et. al. The experimental results over 8 high-dimensional datasets have shown that the proposed method allows saving much execution time than the original one, but retains the equivalent classification performance as well as the equivalent FRBCS complexity.

Keywords: Hedge Algebras, fuzzy classification system, feature selection, high-dimensional dataset.

1. INTRODUCTION

The fuzzy rule based classification system (FRBCS) design problem is one of the concerned study trends in the data mining field and has achieved many successful results. The advantage of this model is that the end-users can use the high interpretability fuzzy rule based knowledge extracted automatically from numerical data as their knowledge.

In the fuzzy set theory approaches for designing FRBCS [1 - 4], the fuzzy sets used to design the fuzzy partitions are pre-specified and the linguistic labels are intuitively assigned to the fuzzy sets, so there is not any constraint between the linguistic terms and their fuzzy sets. When necessary, a genetic fuzzy system is developed to adjust the fuzzy set parameters to achieve the optimal fuzzy partitions. Due to the separation between the term-meaning and their fuzzy sets, the fuzzy sets are deformed after the learning processes. Therefore, it affects the interpretability of the fuzzy rule based systems of the classifiers.

Hedge algebras (HAs) [5-9] take advantage of the algebraic approach that allows to model and design the linguistic terms integrated with their fuzzy sets for FRBCSs. It exploits the inherent semantic order of the linguistic terms allows to generate the semantic constraints between the terms and their integrated fuzzy sets. Based on this formalism, a method to design genetically linguistic terms along with their integrated triangular fuzzy sets to construct an effective fuzzy rule based classifier has been introduced in [10]. To answer the question if trapezoidal fuzzy sets can be used instead of triangular fuzzy sets in the above design method, the so-called enlarged hedge algebras (EnHAs) have been developed in [11], in which the concept of the semantic core of words was introduced. As fuzzy sets, the core of the trapezoids are interval-cores, which can present the core of the term semantics as the numeric values. The computer simulations have shown that the use of trapezoids outperforms the use of triangles in both the ordinary HAs based methodology and the fuzzy set approach.

The time consuming of most of the FRBCS design method is the fuzzy rule generation processes. With the FRBCS design method based on HAs methodology, each feature space is partitioned to *k*-similarity fuzzy intervals, thus, all similarity fuzzy intervals of all features define the hypercubes. From each hypercube containing a data pattern, a fuzzy rule with the length *n* is generated, where *n* is the number of features. The total of this type of rule is $|D|$, where $|D|$ is the number of data patterns. To generate all fuzzy rules with the length from 1 to L less than *n*, a set of fuzzy combinations must be generated. The number of fuzzy combinations is $\sum_{i=1}^{L} C_n^i$, leading to the maximum number of the generated candidate fuzzy rules is $|D| \times \sum_{i=1}^{L} C_n^i$. The candidate fuzzy rules are obtained after removing the inconsistent rules having identical antecedents but different consequence classes. The cardinality of the candidate fuzzy rule set depends on the data distributions and it is still quite high after removing the inconsistent rules. Thereby, the number of candidate fuzzy rules generated by the FRBCS design method based on HAs methodology does not depend on the number of used linguistic terms but still depends on the number of dataset features. Therefore, the main drawback of the FRBCS design method proposed in [11] which limits its application to the high-dimensional datasets is that the number of fuzzy combinations grows with the increase of the dataset features leading to the number of candidate fuzzy rules extensively increases. Ex., the maximum number of the generated fuzzy combinations is **36,050** and the maximum number of the generated candidate fuzzy rules is **7,498,400** for the Sonar dataset (see section 4) with $n = 60$, $|D| = 208$ and $L = 3$. The number of fuzzy combinations is quite high, thus leading to a slow-running of the fuzzy rule generation process. Therefore, a quite good technique [12-15] needed to be applied to reduce a large amount of fuzzy combinations, but also tries to retain a suitable classification performance. For the example above, if the number of features is reduced to **9**, by making all possible combinations, the number of fuzzy combinations is only **129**, the number of generated fuzzy rules is **26,832** and after removing the inconsistent rules, the number of generated candidate fuzzy rules is **15,482**. From the analysis above, the application of an feature reduction method for the high-dimensional datasets needs to be taken into account.

To reduce the running time of the fuzzy rule generation processes, a steady-state genetic algorithm for extracting fuzzy classification rules from data (SGERD) proposed in [12] is applied to the FRBCS design method based on HAs methodology in [13]. The SGERD algorithm shows the efficiency of reducing the rule generation time and has a good scalability when applied to deal with the high-dimensional problems. Howerver, as shown in [14], this method is not good in comparison with the other methods in Friedman's test with the results obtained in the test data.

This paper presents an approach to reduce a large amount of dataset features to tackle the high-dimensional dataset problem for the method proposed in [11] by utilizing the feature selection technique using dynamic weights proposed in [15]. Feature selection is a technique to select a small subset of relevant features having the most discriminating information from the set of original features because the data contain many redundant features. The advantage of this feature selection technique is that it does not only eliminate redundant features and select the most relevant ones, but also tries to retain useful intrinsic feature groups. By using two fundamental information theory concepts, mutual information (MI) and conditional mutual information (CMI), a new scheme for feature relevance, interdependence and redundancy analysis has been introduced [15].

For the proposed method in this paper, the continuous valued features are partitioned into a particular number of clusters by applying the fuzzy c-means clustering technique together with the PBMF cluster validity index function [15, 16] instead of discretizing them into multiple intervals using MDL supervised discretization method [17] used in [15].

The rest of this paper is organized as follows: Section 2 is a short brief description of the FRBCS design based on the EnHAs. Section 3 presents the application of a feature selection technique for the FRBCS design based on the EnHAs. Section 4 represents our experimental results and discussion. Concluding remarks are included in Section 6.

2. FUZZY RULE BASED CLASSIFIER DESIGN BASED ON THE ENLARED HEDGE ALGEBRAS

The fuzzy rule based knowledge of FRBCS used in this paper is the weighted fuzzy rules in the following form [4, 10, 11]:

Rule R_q: IF $\mathcal{X}_1$ is $A_{q,1}$ AND ... AND $\mathcal{X}_n$ is $A_{q,n}$ THEN C_q with CF_q, for $q=1, \ldots, N$ (1)

where $\mathfrak{X} = \{\mathcal{X}_j, j = 1, .., n\}$ is a set of n linguistic variables corresponding to n features of the dataset $\boldsymbol{D}$, $A_{q,j}$ is the linguistic terms of the j^{th} feature F_j, C_q is a class label, each dataset includes M class labels, and CF_q is the weight of rule R_q. The rule R_q can be written as the following short form:

$$A_q \Rightarrow C_q \text{ with } CF_q, \text{ for } q=1, \ldots, N \tag{2}$$

where $\boldsymbol{A}_q$ is the antecedent part of the q^{th}-rule.

A FRBCS design problem $\mathcal{P}$ is defined as: a set $\boldsymbol{P} = \{(\boldsymbol{d}_p, C_p) \mid \boldsymbol{d}_p \in \boldsymbol{D}, C_p \in \boldsymbol{C}, p = 1, \ldots, m;\}$ of m patterns, where $\boldsymbol{d}_p = [d_{p,1}, d_{p,2}, ..., d_{p,n}]$ is the row p^{th} of n data patterns, $\boldsymbol{C} = \{C_s \mid s = 1, \ldots, M\}$ is the set of M class labels.

Solving the problem $\mathcal{P}$ is to extract from $\boldsymbol{P}$ a set $\boldsymbol{S}$ of fuzzy rules in the form (1) such as to achieve a FRBCS based on $\boldsymbol{S}$ comes with high performance, interpretability and comprehensibility. The FRBCS design method based on the enlarged hedge algebras comprises two following phases [11]:

(1) Design automatically the optimal linguistic terms along with their fuzzy-set-based semantics (trapezoidal fuzzy sets) for each dataset feature by applying an evolutionary

multi-objective optimization algorithm in such a way that its outputs are the consequences of the interacting between the semantics of the linguistic terms and the data.

(2) Extract the optimal fuzzy rule set for FRBCS from the dataset in such a way as to achieve their suitable interpretability–accuracy tradeoff based on the optimal linguistic terms provided by the first phase.

In order to realize two phases mentioned above, each j^{th} feature of a specific dataset is associated with an enlarged hedge algebras AX_j^{en}. With the pre-specified values of $\boldsymbol{\Pi}$, comprising the fuzziness measure $fm_j(c^-)$ of the primary term c^-, the fuzziness measure $\mu(h_{j,i})$ of the hedges and a positive integer k_j for limiting the designed term lengths of j^{th} feature, the fuzziness intervals $\Im_k(x_{j,i}), x_{j,i} \in X_{j,k}$ for all $k \leq k_j$ and the interval quantifying mapping values $f(x_{j,i})$ are computed. By utilizing the generated values $\Im_k(x_{j,i})$ and $f(x_{j,i})$, the trapezoidal-fuzzy-set-based semantics of the terms $X_{j,(kj)}$ are computationally constructed. The set of terms $X_{j,(kj)}$ is the union of the subsets $X_{j,k}$, $k = 1$ to k_j, and the k_j-intervals $\Im_{k_j}(X_{j,i})$ of the terms in each $X_{j,kj}$ constitute a binary partition of the feature reference space. For example, the trapezoidal fuzzy sets of terms with $k_j = 2$ is denoted in Figure 1.

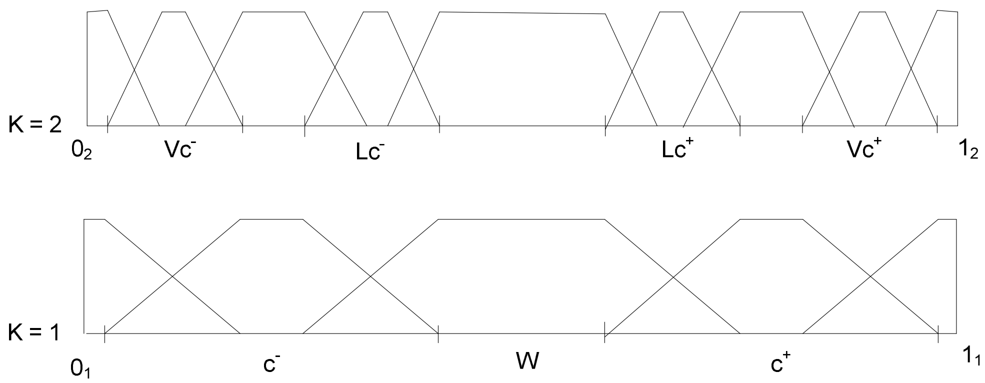

Figure 1. The trapezoidal fuzzy sets of terms in case of $k_j = 2$.

After the binary partitions of all dataset features are constructed, the next step is to generate fuzzy rules from the dataset $\boldsymbol{P}$. With a specific binary partition at k_j level, there is a unique fuzziness interval $\Im_{k_j}(X_{j,i(i)})$ compatible with the term $x_{j,i(j)}$ containing j^{th}-component $d_{j,l}$ of d_l pattern. All k_j-intervals which contain $d_{j,l}$ component defines a hyper-cube $\boldsymbol{\mathcal{H}}_l$, and fuzzy rules are only induced from this type of hyper-cube. So a *basic fuzzy rule* for the class C_l of p_l is generated from $\boldsymbol{\mathcal{H}}_l$ in the following form:

$$\text{IF } X_1 \text{ is } x_{1,i(1)} \text{ AND } \dots \text{ AND } X_n \text{ is } x_{n,i(n)} \text{ THEN } C_l \qquad (R_b)$$

Each data pattern generates only one basic fuzzy rule with the length n. To generate the fuzzy rule with the length $L \leq n$, so-called the *secondary rules*, some techniques should be used for generating fuzzy combinations, ex., generate all possible combinations or use search tree [14].

$$\text{IF } X_{j_1} \text{ is } x_{j_1,i(j_1)} \text{ AND } \dots \text{ AND } X_{j_t} \text{ is } x_{j_t,i(j_t)} \text{ THEN } C_q \qquad (R_{snd})$$

where $1 \leq j_1 \leq \ldots \leq j_t \leq n$. The consequence class C_q of the rule R_q is determined by the confidence measure $(\boldsymbol{A}_q \Rightarrow C_h)$ of R_q:

$$C_q = argmax\{c(\boldsymbol{A}_q \Rightarrow C_h) | h = 1, \ldots, M\} \tag{3}$$

The confidence measure is computed as:

$$c(\boldsymbol{A}_q \Rightarrow C_h) = \sum_{\boldsymbol{d}_p \in C_h} \mu_{\boldsymbol{A}_q}(\boldsymbol{d}_p) / \sum_{p=1}^{m} \mu_{\boldsymbol{A}_q}(\boldsymbol{d}_p) \tag{4}$$

where $\mu_{\boldsymbol{A}_q}(\boldsymbol{d}_p)$ is the burning of pattern $\boldsymbol{d}_p$ for R_q and commonly computed as:

$$\mu_{A_q}(\boldsymbol{d}_p) = \prod_{j=1}^{n} \mu_{q,j}(\boldsymbol{d}_{p,j}). \tag{5}$$

The maximum of number fuzzy combinations is $\sum_i^L C_n^i$, so the maximum of the *secondary rules* is $m \times \sum_i^L C_n^i$.

There may be inconsistent rules which have the identical antecedents, but different consequence classes generated from $\boldsymbol{P}$. They are eliminated by confident measure and the rest of rules are called the *candidate fuzzy rules*. To eliminate the less important rules, a screening criterion is used to select a subset S_0 with NR_0 fuzzy rules from the candidate rule set, called an *initial fuzzy rule set*. This process is done by a so-called initial fuzzy rule set generation procedure IFRG($\boldsymbol{Л}$, P, NR_0, L) [4, 10], where $\boldsymbol{Л}$ is a set of the semantic parameter values and L is the maximum of rule length.

The different given values of the semantic parameters will generate the different binary partition of the feature reference space leading to the different classification performance of a specific dataset. Therefore, in order to get the best ones for a specific dataset, an evolutionary algorithm is used to find the optimal semantic parameter values for generating S_0. The number of the initial fuzzy rules NR_0 is quite large, so an evolutionary algorithm is implemented to find the expected optimal solution. For more details, see [10, 18].

3. AN APPLICATION OF A FEATURE SELECTION TECHNIQUE FOR THE FRBCS DESIGN BASED ON THE ENALRGED HEDGE ALGEBRAS

3.1. Some Concepts of Information Theory

This subsection presents a short brief description of some basic concepts of information theory [15]: entropy and mutual information used to measure the uncertainty of random variables and the information shared by them. Suppose X is a discrete random variable, the entropy $H(X)$ of X is defined as:

$$H(X) = -\sum_{x \in X} p(x) \log(p(x)). \tag{6}$$

where $p(x) = Pr(X = x)$ is the probability distribution function of X.

X and Y is a pair of discrete random variables, the joint entropy $H(X, Y)$ is defined as:

$$H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log(p(x, y)) \tag{7}$$

where $p(x, y)$ is a joint probability distribution which models the relationships between the variables.

When the entropy of the variable X conditioned on the variable Y, we have the conditional entropy $H(X|Y)$ defined as:

$$H(X|Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log(p(x|y)) \tag{8}$$

Mutual information (MI) of two random variables X and Y is a measure of their mutual dependence and is defined as:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log(\frac{p(x,y)}{p(x)p(y)}) \tag{9}$$

The above expression can be re-expressed in terms of joint and conditional entropies, so it is equivalent to as the following:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{10}$$

Thus, the MI between X and Y can be interpreted as the reduction in uncertainty about X after observing Y.

Conditional mutual information (CMI) is defined as the amount of information shared by variables X and Y, when Z is known. It is formally defined by:

$$I(X;Y|Z) = \sum_{z \in Z} \sum_{y \in Y} \sum_{x \in X} p(x,y,z) \log(\frac{p(z)p(x,y,z)}{p(x,z)p(y,z)}) \tag{11}$$

CMI can also be interpreted as the reduction in the uncertainty of X due to Y when Z is known.

3.2. Feature Selection Technique Using Dynamic Weights

Feature selection is a way helps to reduce a large amount of dataset features by selecting a small subset of relevant features from the set of the original ones in order to improve the performance of the learning algorithms. This subsection presents the feature technique using dynamic weight proposed in [15]. This technique does not only eliminate redundant features which are highly correlated with the selected ones as other techniques, but also consider interdependent features which are weak as individuals, but have strong discriminatory power as a group by introducing a new scheme for feature relevance, interdependence and redundancy analyses.

Relevance analysis is used to overcome the drawback of mutual information which tends to favor features with more values by using the symmetrical measure and it is defined as:

$$U(X,Y) = 2 \times \frac{I(X;Y)}{H(X)+H(Y)} \quad (0 \le U(X,Y) \le 1) \tag{12}$$

The redundancy and the interdependence of the candidate features are evaluated by combining MI and CMI. A feature which has one or more other features correlated with is considered to be redundant and the relevance of it to the target class can be reduced by the knowledge of any one of the correlated features. Thus, a feature f_i is considered to be redundant with the feature f_j if the hereafter in-equation is satisfied:

$$I(f_i; class|f_j) \le I(f_i; class) \tag{13}$$

The relative Redundancy Ratio between two features $RR(i, j)$ represents the reduction ratio of relevance between the feature f_i and the target class due to the feature f_j and is defined as:

$$RR(i,j) = 2 \times \frac{I(f_i;class|f_j) - I(f_i;class)}{H(f_i)+H(class)} \quad (-1 \le RR(i,j) \le 0) \tag{14}$$

Two features f_i and f_j are interdependent on each other if the hereafter in-equation is satisfied:

$$I(f_i; class|f_j) \ge I(f_i; class) \tag{15}$$

The interdependent ratio $IR(i, j)$ between f_i and f_j which denotes the increase's ratio of relevance between f_i and the target class by new feature joining is defined as:

$$IR(i,j) = 2 \times \frac{I(f_i;class|f_j) - I(f_i;class)}{H(f_i) + H(class)} \quad (0 \le IR(i,j) \le 1) \tag{16}$$

Both $RR(i, j)$ and $IR(i, j)$ are unified as correlation ratio $CR(i, j)$:

$$CR(i,j) = \begin{cases} IR(i,j), I(f_i;class|f_j) > I(f_i;class) \\ RR(i,j), I(f_i;class|f_j) \le I(f_i;class) \end{cases} \tag{17}$$

It is obviously that $-1 \le CR(i,j) \le 1$.

Based on the above information metrics, a dynamic weighting-based feature selection algorithm for ranking features, abbreviated as DWFS, has been proposed in [15]. Hereafter is the pseudo code of the algorithm described in details:

Algorithm 1. DWFS: the adapted algorithm proposed in [15].
Input: A training sample ***D*** with feature space ***F*** and the target ***C***.
Output: The subset *S* selected from δ features
Initialize parameters: $k = 1$, $S = \emptyset$;
Initialize the weight $w(f)$ for each feature f in F to 1 equally;
Calculate the value of $U(f$, class) for each feature f in F;
While $k \le \delta$ **do**
  For each candidate feature $f \in F$ **do**
    Calculate $J(f) = R(f, class) \times w(f)$;
  End;
  Choose the candidate feature f_j with the largest $J(f)$;
  Add f into the selected subset $S = S \cup \{f_j\}$;
  $F = F \setminus \{f_j\}$;
  For each candidate feature $i \in F$ **do**
    Calculate the Correlation ratio $CR(i, j)$;
    Update $w(i)$ by $w(i) = w(i) \times (1 + CR(i,j))$;
  End;
  $k = k + 1$;
End.

The complexity of DWFS algorithm is $O(n \times \delta)$ as already proofed in [15], where, *n* is the number of original features and δ is the number of selected features.

3.3. The Application of the DWFS for the FRBCS Design Based on the EnHAs

The FRBCS design based on the enlarged hedge algebras methodology proposed in [11] is an efficient way to extract the fuzzy rule based systems from a given numerical dataset for the fuzzy rule based classifier. However, as described in the first section, dealing with the high-dimensional datasets is still a critical issue needed to be considered. This subsection presents an approach to tackle the high-dimensional dataset issue for the FRBCS design based on the enlarged hedge algebras by utilizing the DWFS algorithm described in the previous subsection.

Hence, the extended method proposed in this paper comprises three phases by inserting the feature selection preprocessing mechanism into the original method as the first phase:

(1) For a given dataset, the continuous valued features are partitioned into a particular number of clusters by applying the fuzzy c-means clustering technique together with the PBMF cluster validity index function [16, 19] and then apply the DWFS algorithm to select a subset of the most discriminating features.

(2) Design automatically the optimal linguistic terms along with their fuzzy-set-based semantics (trapezoidal fuzzy sets) for each feature of the subset of the dataset having only the features selected by the first phase, so-called the selected training set.

(3) Extract the optimal fuzzy rule set for the FRBCS from the selected training set.

In the first phase, the continuous valued features are clustered by the fuzzy c-means clustering technique. After the clustering process, the real-valued data is partitioned into $v > 0$ clusters produced by the process and each cluster is assigned a sequence number in order to achieve the discrete values of the processed feature.

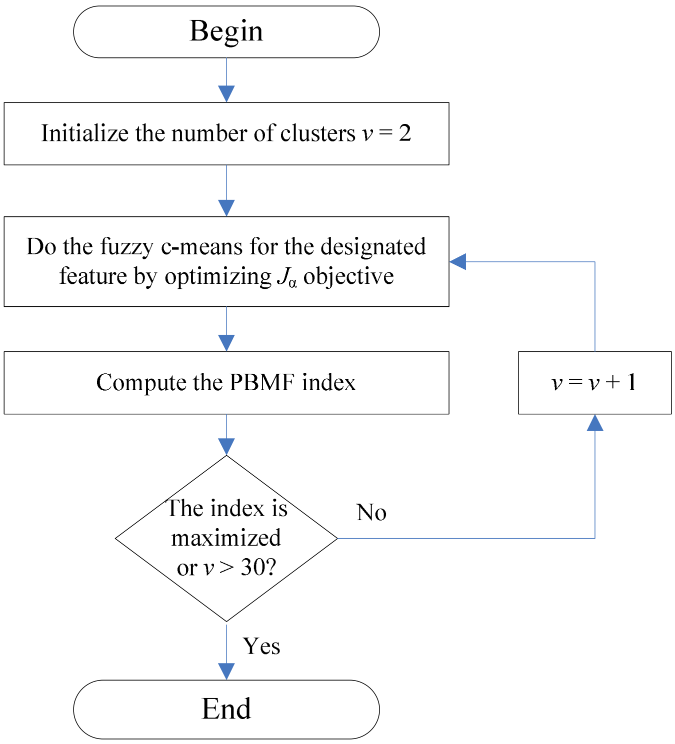

Figure 2. The flow chart of the fuzzy c-means clustering technique together with the PBMF index validation.

Let $Y = \{y_1, \ldots, y_m\}$ be the dataset of j^{th}-feature. Fuzzy c-means clustering technique optimizes the following objective function:

$$J_\alpha = \sum_{i=1}^{m} \sum_{j=1}^{v} \mu_{i,j}^{\alpha} \left\| y_i - v_j \right\|^2, 1 < \alpha < \infty, \tag{18}$$

where v is the number of clusters, $\mu_{i,j}$ is the membership degree of y_i in the cluster j, VJ is the centroid of the cluster, $\alpha > 1$ is the fuzzifier exponent which make the partitions more or less fuzzy. The membership degree $\mu_{i,j}$ and the cluster centroid v_j updated by the optimization process:

$$\mu_{i,j} = \frac{1}{\sum_{k=1}^{v}\left(\frac{\|y_i - v_j\|}{\|y_i - v_k\|}\right)^{\frac{2}{\propto - 1}}} \tag{19}$$

$$v_j = \frac{\sum_{i=1}^{m}\mu_{i,j}^{\alpha}\times y_i}{\sum_{i=1}^{m}\mu_{i,j}^{\alpha}} \tag{20}$$

The optimization process stops when the number of iterations reaches the maximum number or $|J_{\alpha}^{(k+1)} - J_{\alpha}^{k}| < \varepsilon$, where $0 < \varepsilon < 1$ and k is the current number of iterations.

The PBMF index method [16, 19] is used for optimizing the number of clusters and it is defined as:

$$V_{PBMF} = \left(\frac{1}{v}\times\frac{E_1}{J_\alpha}\times Z_v\right)^2 \tag{21}$$

where $E_1 = \sum_{j=1}^{m}\|y_j - e\|$ with e is the dataset's centroid and $Z_v = max_{i,j=1}^{v}\|v_i - v_j\|$.

The flow chart of the fuzzy c-means clustering technique together with the PBMF index validation is denoted in Figure 2.

After the clustering processes, all real-valued features are discretized for the input of the feature selection process using the DWFS algorithm described above.

The two last phases are the two phases of the FRBCS design based on the enlarged hedge algebras proposed in [11], except the training set is the selected set instead of the original one.

4. EXPERIMENTAL RESULTS AND DISCUSSION

This section presents the experimental results of applying the feature selection technique described in the above sections as a preprocessing method to the FRBCS design based on the enlarged hedge algebras methodology proposed in [11] in comparison with the original method over some real world high-dimensional datasets that can be found on the KEEL-Dataset repository: http://sci2s.ugr.es/keel/datasets.php. All the implementations for validating have been implemented using C#, and all the experiments have been performed using an Intel Core i3-550, 3.2GHz CPU with 2 GB of memory and running Microsoft Windows XP 32-bit. The 8 high dimensional datasets used to validate in this study are listed in the Table 1.

Table 1. The high dimensional datasets used in this study.

No.	Dataset name	Number of attributes	Number of classes	Number of patterns
1	Bands	19	2	365
2	Dermatology	34	6	358
3	Hepatitis	19	2	80
4	Ionosphere	34	2	351
5	Sonar	60	2	208
6	Spambase	57	2	4597
7	Spectfheart	44	2	267
8	Wdbc	30	2	569

Table 2. The number of selected features of the validated datasets.

No.	Dataset name	Number of attributes	S_n	S_{2n}
1	Bands	19	6	8
2	Dermatology	34	7	10
3	Hepatitis	19	6	8
4	Ionosphere	34	7	10
5	Sonar	60	9	12
6	Spambase	57	9	12
7	Spectfheart	44	8	11
8	Wdbc	30	7	9

First of all, the feature selection preprocessing technique is applied to each dataset to select the most discriminating feature subset. Two feature's quantities of $\lceil\sqrt{n}\rceil + 1$ and $\lceil\sqrt{2n}\rceil + 1$ are used to validate, where n is the number of the original dataset, for convenience, named as S_n and S_{2n} respectively. The feature's quantity of the original dataset is named as N. After this phase, the number selected features of the validated datasets are listed in the Table 2.

The subsets of data with the selected features of the corresponding validated datasets after applying the feature selection preprocessing are taken into account. The same *ten-folds cross validation* method is applied to every subset of the validated datasets and the original ones, i.e., each of them is randomly partitioned into 10 folds, 9 folds for the training phase and one fold for the testing phase. Three trials of the FRBCS design method based on HAs are executed for each of ten folds and, hence, it permits to extract 30 (= 3 times × 10 folds) FRBCSs from the data.

To limit the searching space in the learning process, the same constraints on the semantic parameter values is applied as examined in [11]. i.e., we have: the number of both negative hedge and positive hedge is 1, and assume that the negative hedge is L and the positive hedge is V; 0.00001 $\leq$ $fm(\mathbf{0})$, $fm(\mathbf{1})$ $\leq$ 0.01; $0.2 \leq fm\left(c^{-}\right) \leq 0.6$; 0.0001 $\leq$ $fm(W)$ $\leq$ 0.2; $0.2 \leq \mu\left(L\right) \leq 0.6$; $0.0001 \leq \mu\left(h_0\right) \leq 0.5$ and $1 \leq k_j \leq 3$.

The optimization algorithm used in this study is the multi-objective particle swarm optimization with fitness sharing proposed in [20]. It is an efficient algorithm as presented in [18].

The semantic parameter optimization process [11] has been run with the following parameters: the number of generations = 250, the same as examined in [11]; the number of particles of each generation = 600; Inertia coefficient = 0.4; the self cognitive factor = 0.2; the social cognitive factor = 0.2; the number of initial fuzzy rules is equal to the number of attributes; the maximum of rule length is 1.

The fuzzy rule selection process [11] has been run with the same parameters of the semantic parameter optimization process, except the number of generations = 1000; the number of particles of each generation = 600; the number of initial fuzzy rules $|\mathbf{S}_0|$ = *300 × number of classes*; the maximum of rule length = 3.

The running time in the *hh:mm:ss* format of the initial fuzzy rule generation processes from the validated datasets with and without applying the feature selection preprocessing are listed in the Table 3, where noted that *L2* and *L3* are the running times in case the maximum of fuzzy rule length is 2 and 3, respectively.

Table 3. The comparison of the running times of the initial fuzzy rule generation processes.

No.	Dataset name	*N*		S_n		S_{2n}	
		L2	***L3***	***L2***	***L3***	***L2***	***L3***
1	Bands	00:00:18	00:22:45	00:00:00	00:00:01	00:00:00	00:00:04
2	Dermatology	00:02:54	09:17:00	00:00:00	00:00:00	00:00:00	00:00:07
3	Hepatitis	00:00:02	00:01:12	00:00:00	00:00:00	00:00:00	00:00:00
4	Ionosphere	00:13:43	38:34:11	00:00:00	00:00:03	00:00:03	00:00:31
5	Sonar	01:53:48	-	00:00:01	00:00:08	00:00:04	00:01:23
6	Spambase	04:05:01	-	00:00:11	00:01:25	00:00:29	00:13:45
7	Spectfheart	00:11:27	66:12:07	00:00:00	00:00:03	00:00:01	00:00:28
8	Wdbc	00:07:16	10:37:12	00:00:00	00:00:02	00:00:00	00:00:15

As shown in the Table 3, the running time of the initial fuzzy rule generation processes after applying the feature selection to the original datasets are reduced very much, especially, in case the fuzzy rule length is 3 (in case of *L3* as in the Table 3). Ex., the initial fuzzy rule extraction time from the original Dermatology dataset in case of *L3* is 09:17:00 or 33,420 seconds, which is greater than 33,420 and 4,774 times in case of the feature's quantities of $\lceil\sqrt{n}\rceil + 1$ (0 seconds) and $\lceil\sqrt{2n}\rceil + 1$ (07 seconds) respectively. The "-" values mean that the fuzzy rule generation processes are too slow that the results cannot be obtained. That while we usually limit the maximum of rule length to 2 with the datasets having the number of features greater than and equal to 30 in the previous studies.

The experimental results of the classification performance of the application of the feature selection technique presented in the above section for the FRBCS design are shown in the Table 4, where note that *#R*, *#C* and *#R*#C* are the number of fuzzy rules, the number of conditions and the complexity of the extracted fuzzy rule set respectively; P_{tr} ≔ the performance in the training phase and P_{te} ≔ the performance in the testing phase; The ≠*C* and ≠*Pte* columns represent the differences of the complexities and the performances of the compared methods respectively. Specifically, the average results of the three validated methods are not much different. Therefore, the final conclusion should rely upon the statistic studies given in the Table 5 and the Table 6 in which the Wilcoxon's signed-rank tests [21] have been applied to test the complexities and performances of the fuzzy rule bases extracted by three methods respectively. It is assumed that the two compared versions are statistically equivalent (null-hypothesis).

Table 4. The comparison of the classification performances of the original datasets and their subsets of $\lceil\sqrt{2n}\rceil + 1$ and $\lceil\sqrt{n}\rceil + 1$ features.

No.	Dataset name	*N*			S_n			≠*C*	≠*Pte*	S_{2n}			≠*C*	≠*Pte*
		#R*#C	P_{tr}	P_{te}	***#R*#C***	P_{tr}	P_{te}			***#R*#C***	P_{tr}	P_{te}		
1	Bands	58.20	78.19	73.46	51.78	73.05	70.52	*6.42*	*2.94*	52.36	73.07	70.35	*5.84*	*3.11*
2	Dermato.	182.84	96.37	94.40	269.04	90.37	89.18	*-86.20*	*5.22*	328.91	95.94	94.14	*-146.07*	*0.26*

3	Hepatitis	25.53	93.68	89.28	20.52	93.52	88.51	*5.01*	*0.77*	23.32	95.81	**89.60**	*2.21*	***-0.32***
4	Ionosphere	88.03	94.69	91.56	81.75	93.74	**91.65**	*6.28*	***-0.09***	76.04	94.84	**92.98**	*11.98*	***-1.42***
5	Sonar	49.31	87.59	78.61	41.61	86.96	**79.66**	*7.70*	***-1.05***	49.98	89.39	**81.79**	*-0.67*	***-3.18***
6	Spambase	17.28	85.62	84.94	30.28	87.52	**86.93**	*-13.00*	***-1.99***	36.00	87.68	**87.01**	*-18.72*	***-2.07***
7	Spectfheart	22.07	82.06	79.42	21.32	83.59	**81.55**	*0.75*	***-2.13***	25.32	84.74	**82.55**	*-3.25*	***-3.13***
8	Wdbc	25.04	97.08	96.78	31.15	97.12	96.20	*-6.11*	*0.58*	29.15	97.06	96.43	*-4.11*	*0.35*
Mean		**58.54**	**89.41**	**86.06**	**68.43**	**88.23**	**85.53**			**77.64**	**89.82**	**86.86**		

The abbreviation terms used in the Table 5 and 6: VS column is the list of the name of the method which we want to compare with; E. is Exact; A. is Asymptotic; Inte. is Interval and Conf. is Confidence.

As shown in the Table 5, the complexities of the FRBCSs extracted from the original datasets (n features) are compared with the complexities of those extracted from the datasets with the subsets of selected features in both cases of the feature's quantities of $\lfloor\sqrt{n}\rfloor + 1$ and $\lfloor\sqrt{2n}\rfloor + 1$ using the Wilcoxon's signed-rank test at level $\alpha = 0.05$. Since all R^- values which are the sum of the ranking results of the FRBCSs extracted from the original datasets are greater than the critical value of T Wilcoxon distribution [22] associated with the number of datasets $N_{ds} = 8$ and $p = 0.05$, where the critical value is 5, all the null-hypotheses cannot be rejected. Therefore, we do not need to take the complexity of the FRBCS into account in the comparisons.

Table 5. The comparison result of the fuzzy rule complexities using the Wilcoxon's signed rank test at level α = 0.05.

VS	R^+	R^-	E. *P*-value	A. *P*-value	Conf. Inte.	Exact. Conf.	Hypothesis
S_{2n}	23	13	≥ 0.2	0.441209	[-74.66 , 5.84]	0.96094	Not rejected
S_n	18	18	≥ 0.2	0.944183	[-42.725 , 6.42]	0.96094	Not rejected

The comparison of the extracted FRBCS performances using Wilcoxon's signed-rank test at level α = 0.05 is shown in the Table 6. All the null-hypotheses cannot be rejected, so we can state that both the feature's quantities of $\lfloor\sqrt{n}\rfloor + 1$ and $\lfloor\sqrt{2n}\rfloor + 1$ do not affect the classification performance of the FRBCS design based on the enlarged hedge algebras methodology. To reduce the running time of the fuzzy rule generation process of the FRBCS design based on the enlarged hedge algebras methodology for the high dimensional datasets, the proposed feature selection preprocessing should be applied.

Table 6. The comparison result of the fuzzy rule based classification performances using the Wilcoxon's signed rank test at level $\alpha = 0.05$.

VS	R^+	R^-	E. *P*-value	A. *P*-value	Conf. Inte.	Exact. Conf.	Hypothesis
S_{2n}	10	26	≥ 0.2	1	[-2.625 , 1.395]	0.96094	Not rejected
S_n	20	16	≥ 0.2	0.726286	[-1.59 , 2.94]	0.96094	Not rejected

5. CONCLUSION

This paper presents an application of a feature selection technique as the preprocessing mechanism for the fuzzy rule based classifier design based on the enlarged hedge algebras methodology for the high-dimensional datasets. By utilizing this technique, the extended method for the fuzzy rule based classifier design based on the enlarged hedge algebras has been proposed to tackle the high-dimensional datasets comprising three phases by inserting the feature selection preprocessing mechanism into the original method as the first phase. The experimental results over 8 high-dimensional datasets have shown that the proposed method allows saving much execution time than the original one, but retains the equivalent classification performance as well as the equivalent FRBCS complexity.

Acknowledgement. This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.05-2013.34.

REFERENCES

1. Alcalá R., Nojima Y., Herrera F., Ishibuchi H. - Generating single granularity-based fuzzy classification rules for multi-objective genetic fuzzy rule selection, in: Proceedings of the 2009 IEEE International Conference on Fuzzy Systems, Jeju (Korea), 2009, pp. 1718–1723.
2. Alcalá R., Nojima Y., Herrera F., Ishibuchi H. - Multi-objective genetic fuzzy rule selection of single granularity-based fuzzy classification rules and its interaction with the lateral tuning of membership functions, Journal Soft Computing **15** (12) (2011) 2303–2318.
3. Chen Ji-lin, Hou Yuan-long, Xing Zong-y, Jia Li-min, Tong Zhong-zhi - A Multi-objective Genetic-based Method for Design Fuzzy Classification Systems, IJCSNS International Journal of Computer Science and Network Security **6** (8A) (2006) 110-118.
4. Ishibuchi H., Yamamoto T. - Fuzzy Rule Selection by Multi-Objective Genetic Local Search Algorithms and Rule Evaluation Measures in Data Mining, Fuzzy Sets and Systems **141** (1) (2004) 59-88.
5. Ho N. C., Wechler W. - Hedge algebras: an algebraic approach to structures of sets of linguistic domains of linguistic truth variables, Fuzzy Sets and Systems **35** (3) (1990) pp. 281-293.
6. Ho N. C., Wechler W. - Extended hedge algebras and their application to fuzzy logic, Fuzzy Sets and Systems **52** (1992) 259–281.
7. Ho N. C, Nam H. V., Khang D. T., Le H.C. - Hedge Algebras, Linguistic-valued logic and their application to fuzzy reasoning, Internat, J. Uncertain. Fuzziness Knowledge-Based Systems **7** (4) (1999) 347–361.
8. Ho N. C., Long N. V. - Fuzziness measure on complete hedges algebras and quantifying semantics of terms in linear hedge algebras, Fuzzy Sets and Systems **158** (2007) pp. 452-471.
9. Ho N. C. - A topological completion of refined hedge algebras and a model of fuzziness of linguistic terms and hedges, Fuzzy Sets and Systems **158** (2007) 436–451.

10. Ho N. C., Pedrycz W., Long D. T., Son T. T. - A genetic design of linguistic terms for fuzzy rule based classifiers, International Journal of Approximate Reasoning **54** (1) (2013) 1-21.
11. Ho N. C., Son T. T., Phong P. D. - Modeling of a semantics core of linguistic terms based on an extension of hedge algebra semantics and its application, Knowledge-Based Systems **67** (2014) 244–262.
12. Mansoori E. G., Zolghadri M. J., Katebi S. D. - SGERD: A Steady-State Genetic Algorithm for Extracting Fuzzy Classification Rules From Data **16** (4) (2008) 1061-1071.
13. Ho N. C., Thong H. V., Long N. V. - An evolutionary method to generate fuzzy rule-based system for classification problems with semantics order of language, Journal of Computer Science and Cybernetics **28** (4) (2012) 333–345.
14. Alcalá-Fdez J., Alcalá R., and Herrera F. - A Fuzzy Association Rule-Based Classification Model for High-Dimensional Problems With Genetic Rule Selection and Lateral Tuning, IEEE Transactions on Fuzzy System **19** (5) (2011) 857-872.
15. Sun X., Liu Y., Xu M., Chen H., Han J., Wang K.- Feature selection using dynamic weights for classification, Knowledge-Based Systems **37** (2013) 541-549.
16. Pakhira M. K., Bandyopadhyay S., Maulik U. - A study of some fuzzy cluster validity indices, genetic clustering and application to pixel classification, Fuzzy Sets and Systems **155** (2005) 191–214.
17. Fayyad U. M., Irani K. B. - Multi-interval discretization of continuous-valued attributes for classification learning, In Proceedings of 13th International Joint Conference on Artificial Intelligence, Morgan Kaufmann, Chambéry, France (1993) 1022–1027.
18. Phong P. D., Ho N. C., Thuy N. T. - Multi-objective Particle Swarm Optimization Algorithm and its Application to the Fuzzy Rule Based Classifier Design Problem with the Order Based Semantics of Linguistic Terms, In Proceedings of The 10th IEEE RIVF International Conference on Computing and Communication Technologies (RIVF-2013), Hanoi, Vietnam, 2013, pp. 12-17.
19. Pakhira M.K., Bandyopadhyay S., Maulik U. - Validity index for crisp and fuzzy clusters, Pattern Recognition **37** (2004) 487–501.
20. Maximino S. L. - Multi-Objective Optimization using Sharing in Swarm Optimization Algorithms, Doctor thesis, School of Computer Science, The University of Birmingham, 2006.
21. Demˇsar J. - Statistical Comparisons of Classifiers over Multiple Data Sets, Journal of Machine Learning Research **7** (2006) 1–30.
22. Zar J. - Biostatistical Analysis, Prentice-Hall, Upper Saddle River, NJ, pp. 758–760.

TÓM TẮT

AN APPLICATION OF FEATURE SELECTION FOR THE FUZZY RULE BASED CLASSIFIER DESIGN BASED ON AN ENLARGED HEDGE ALGEBRAS FOR HIGH-DIMENSIONAL DATASETS

Pham Dinh Phong[1,2]

[1]*Prévoir Vietnam, 23 Phan Chu Trinh, Hanoi, Vietnam*

[2]*University of Engineering and Technology, Hanoi National University, 144 Xuan Thuy, Cau Giay, Hanoi, Vietnam*

Email: *dinhphongpham@gmail.com*

Các phương pháp thiết kế hệ phân lớp dựa trên luật mờ dạng if-then đã và đang được nghiên cứu rộng rãi trong những năm gần đây. Một trong các phương pháp thiết kế hệ phân lớp dựa trên luật mờ xuất sắc sử dụng phương pháp luận đại số giá tử mở rộng làm cơ chế hình thức cho việc thiết kế tối ưu các từ ngôn ngữ cùng với ngữ nghĩa dựa trên các tập mờ hình thang của chúng đã được đề xuất bởi nhóm tác giả Nguyễn Cát Hồ. Cũng giống như các tiếp cận khác, một trong những khó khăn cần phải khắc phục đối với tiếp cận này là xử lí các tập dữ liệu mẫu có số chiều lớn và nhiều mẫu dữ liệu. Bài báo trình bày một tiếp cận để giải bài toán phân lớp với tập dữ liệu có số chiều lớn đối với phương pháp thiết kế hệ phân lớp dựa trên luật mờ sử dụng phương pháp luận đại số gia tử mở rộng bằng việc áp dụng giải thuật lựa chọn đặc trưng được đề xuất bởi nhóm tác giả Xin Sun. Kết quả thực nghiệm với 8 tập dữ liệu mẫu có số chiều lớn cho thấy phương pháp được đề xuất cho phép giảm đáng kể thời gian thực thi nhưng vẫn đảm bảo được hiệu suất phân lớp cũng như độ phức tạp của hệ phân lớp thu được.

Từ khóa: đại số gia tử, hệ phân lớp mờ, lựa chọn đặc trưng, tập dữ liệu có số chiều lớn.