*Journal of Computer Science and Cybernetics, V.31, N.2 (2015), 171–184*
DOI: 10.15625/1813-9663/31/2/5025

# AN APPLICATION OF FEATURE SELECTION FOR THE FUZZY RULE BASED CLASSIFIER DESIGN WITH THE ORDER BASED SEMANTICS OF LINGUISTIC TERMS FOR HIGH-DIMENSIONAL DATASETS

PHAM DINH PHONG[1,2]

[1] *Prévoir Vietnam, 23 Phan Chu Trinh, Hanoi, Vietnam*
[2] *Ph.D. student, University of Engineering and Technology, Hanoi National University*

**Abstract.** The fuzzy rule based classification system (FRBCS) design methods, whose fuzzy rules are in the form of if-then sentences, have been under intensive study during last years. One of the outstanding FRBCS design methods utilizing hedge algebras as a mathematical formalism is proposed in [9]. As in other methods, a difficult problem with the high-dimensional and multi-instance datasets needs to be solved. This paper presents an approach to tackle the high-dimensional dataset problem for the hedge algebras based classification method proposed in [9] by utilizing the feature selection algorithm proposed in [20]. The experimental results over eight high-dimensional datasets have shown that the proposed method saves much execution time than the original one, while retaining the equivalent classification performance as well as the equivalent FRBCS complexity. The proposed method is also compared with three classical classification methods based on the statistical and probabilistic approaches showing that it is a robust classifier.

**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 [1,2,12,13], 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) [6–8,10,11] 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 that allows generating the semantic constraints between the terms and their integrated fuzzy sets. By utilizing the values of the semantic parameters which include fuzziness measures, fuzziness intervals of terms, semantically quantifying mappings (SQMs) of the hedge algebras [7,8] and a positive integer to limit the term lengths, denoted by $\boldsymbol{\Pi}$, the triangular fuzzy sets of terms are generated automatically. Based on this, a genetic design of linguistic terms of the fuzzy rule based classifiers is determined [9]. This method comprises two phases: the first is the phase to design the optimal linguistic terms along with their triangular fuzzy set based semantics for each dataset feature by adjusting only the semantic parameter values to find their optimal values using an evolutionary algorithm. The second phase is to extract a near optimal FRBCS having a quite suitable interpretability–accuracy trade-offs from the given dataset with the given optimal semantic parameter values provided by the first phase using an evolutionary algorithm for fuzzy rule selection.

The main drawback of the FRBCS design method proposed in [9] 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, i.e. the maximum number of fuzzy combinations is $\sum_{i}^{L} C_n^i$, and the maximum number of the generated candidate fuzzy rules is $|D| \times \sum_{i}^{L} C_n^i$, where $|D|$ is the number of data patterns, $n$ is the number of features and $L$ is the maximum of rule length. The candidate fuzzy rules are obtained after removing the inconsistent rules having identical antecedents, but different consequence classes and their cardinality depend on the data distributions. 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 [3, 14, 20] needs 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**.

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 [9] by utilizing the feature selection technique using dynamic weights proposed in [20]. 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 is introduced [20].

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 [4] used in [20].

The paper is organized as follows: Section 2 is a short brief description of the FRBCS design based on HAs. Section 3 presents the application of a feature selection technique for the FRBCS design based on HAs. Section 4 represents our experimental results and discussion. Concluding remarks are included in Section 5.

## 2. FUZZY RULE BASED CLASSIFIER DESIGN BASED ON THE HEDGE ALGEBRAS METHODOLOGY

The fuzzy rule based knowledge of FRBCS used in this paper is the weighted fuzzy rules in the following form [9, 12]:

$$\text{Rule } R_q\text{: IF } \mathscr{X}_1 \text{ is } A_{q,1} \text{ AND} \ldots \text{AND } \mathscr{X}_n \text{ is } A_{q,n} \text{ THEN } C_q \text{ with } CF_q, \text{ for } q = 1, \ldots, N \tag{1}$$

where $x = \{\mathscr{X}_j, j = 1, \ldots, n\}$ 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:

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

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

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

Solving the problem $\mathscr{P}$ 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 hedge algebras comprises two following phases [9]:

(1) Design automatically the optimal linguistic terms along with their fuzzy-set-based semantics 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.

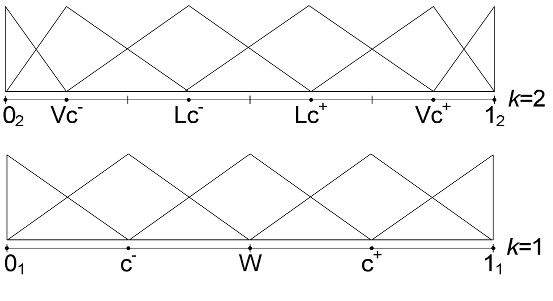

Figure 1: The fuzzy sets of terms in case of $k_j = 2$.

In order to realize two phases mentioned above, each $j^{th}$ feature of a specific dataset is associated with a hedge algebras $AX_j$. 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 $\mathscr{I}_k(x_{j,i})$, $x_{j,i} \in X_{j,k}$ for all $k \le k_j$ and the SQM values $v(x_{j,i})$ are computed. By utilizing the generated values $\mathscr{I}_k(x_{j,i})$ and $v(x_{j,i})$, the fuzzy-set-based semantics of the terms $X_{j,(kj)}$ are computationally constructed. The triangular fuzzy set is commonly assigned to the term $x_{j,i}$. The set of terms $X_{j,(kj)}$ is the union of the subsets $X_{j,k}, k = 1$ to $k_j$, and the $k_j$-similarity intervals $\mathcal{S}_{k_j}(X_{j,i})$ of the terms in each $X_{j,kj+2}$ constitute a binary partition of the feature reference space. For example, the fuzzy sets of terms with $k_j = 2$ is denoted in Figure 1.

After all $k_j$-similarity based 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 $k_j$-similarity interval $\mathcal{S}_{k_j}(X_{j,i(i)})$ compatible with the term $x_{j,i(j)}$ containing $j^{th}$-component $d_{j,l}$ of $d_l$ pattern. All $k_j$-similarity intervals which contain $d_{j,l}$ component defines a hyper-cube $\mathcal{H}_l$, 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 $\mathcal{H}_l$ in the following form:

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

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 [3].

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

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

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

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$(\boldsymbol{\Pi}, \boldsymbol{P}, NR_0, L)$ [9], where $\boldsymbol{\Pi}$ is a set of the semantic parameter values and $L$ is the maximum 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, a multi-objective evolutionary algorithm is used to find the optimal semantic parameter values for generating $\boldsymbol{S}_0$. The objectives of the applied evolutionary algorithm are the classification accuracy of the training set and the average length of the antecedent of fuzzy rule based system. After the applied algorithm produces a set of best semantic parameters $\boldsymbol{\Pi}_{opt}$, any one of the best solutions is taken, $\boldsymbol{\Pi}_{opt,i^*}$, to generate the initial fuzzy rule set

$\boldsymbol{S}_0(\boldsymbol{\Pi}_{opt,i^*})$ which contains $NR_0$ fuzzy rules using IFRG$(\boldsymbol{\Pi}_{opt,i^*}, \boldsymbol{P}, NR_0, \lambda)$. The problem now is to select a subset of fuzzy rules $\boldsymbol{S}$ from $\boldsymbol{S}_0$ satisfying three objectives: the classification accuracy of the training set, the average length of the antecedent and the number of rules of fuzzy rule based system. An evolutionary algorithm is implemented to find the expected optimal solution. For more details, see [9].

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

### 3.1. Some Concepts of Information Theory

This subsection presents a short brief description of some basic concepts of information theory [20]: 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:

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

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:

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

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 $Y$, the conditional entropy $H(X|Y)$ is defined as:

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

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

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

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:

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

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 selection technique using dynamic weight proposed in [20]. This technique does not only eliminate redundant features which are highly correlated with the selected ones as other techniques, but also considers 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:

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

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:

$$I\left(f_i;class|f_j\right)\leq I(f_i;class). \tag{13}$$

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:

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

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

$$I\left(f_i;class|f_j\right)\geq I(f_i;class) \tag{15}$$

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:

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

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

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

It is obviously that $-1\leq CR\left(i,j\right)\leq 1$.

Based on the above information metrics, a dynamic weighting-based feature selection algorithm for ranking features, abbreviated as DWFS, is proposed in [20]. Hereafter is the pseudo code of the algorithm described in details:

```
Algorithm 1. DWFS: the adapted algorithm proposed in [20].
Input: A training sample $\boldsymbol{D}$ with feature space $\boldsymbol{F}$ and the target $\boldsymbol{C}$.
Output: The subset $S$ selected from $\delta$ 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 $k \leq \delta$ 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 $S = S \cup \{f_j\}$;
    $F = F \backslash \{f_j\}$;
    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 [20], where, $n$ is the number of original features and $\delta$ is the number of selected features.

### 3.3. The Application of the DWFS for the FRBCS Design Based on Hedge Algebras

The FRBCS design based on hedge algebras methodology proposed in [9] is an efficient way to extract the fuzzy rule based systems from a given numeric 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 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 [15, 16] 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 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.

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

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

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 partions more or less fuzzy. The membership degree $\mu_{i,j}$ and the cluster centroid $v_j$ updated by the optimization process:

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

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

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 [15, 16] 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 hedge algebras proposed in [9], except the training set is the selected set instead of the original one.

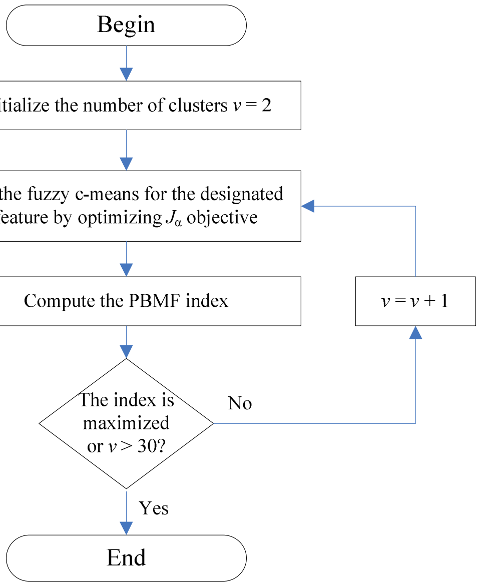
Begin
Initialize the number of clusters $v = 2$
Do the fuzzy c-means for the designated feature by optimizing $J_\alpha$ objective
Compute the PBMF index
$v = v + 1$
The index is maximized or $v > 30$?
No
Yes
End

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

## 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 hedge algebras methodology proposed in [9] 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 i3-3110M, 2.4-GHz CPU with 4 GB of memory and running Microsoft Windows 7 64-bit.The 8 high dimensional datasets used to validate and compare with the original method in this study are listed in the Table 1.

| **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 1: The high dimensional datasets used in this study

| **No.** | **Dataset name** | **Number of attributes** | $\boldsymbol{S}_n$ | $\boldsymbol{S}_{2n}$ |
|---|---|---|---|---|
| 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 |

Table 2: The number of selected features of thevalidated datasets

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 $\sqrt{n}+1$ and $\sqrt{2n}+1$ 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 [9]. 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 $V$; $0 \leq \quad k_j \quad \leq 3$; $0.2 \leq fm_j(c^-) \quad \leq 0.8$; $fm_j(c^-) + fm_j(c^+) = 1$; $0.2 \leq \mu_j(L) \quad \leq 0.8$ and $\mu_j(L) + \mu_j(V) = 1$.

The optimization algorithm used in this study is the multi-objective particle swarm optimization with fitness sharing proposed in [19]. It is an efficient algorithm as presented in [17].

The semantic parameter optimization process [9] has been run with the following parameters: the number of generations = 250, the same as examined in [9]; the number of particles of each generation = 300; 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 [9] 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 $|\mathbf{S}_0| = \mathit{300} \times \mathit{number\ of\ classes}$; 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 time in case the maximum of fuzzy rule length is 2 and 3 respectively.

| **No.** | **Dataset name** | $\boldsymbol{N}$ | | $\boldsymbol{S}_n$ | | $\boldsymbol{S}_{2n}$ | |
|---|---|---|---|---|---|---|---|
| | | ***L2*** | ***L3*** | ***L2*** | ***L3*** | ***L2*** | ***L3*** |
| 1 | Bands | 00:00:20 | 00:20:15 | 00:00:01 | 00:00:02 | 00:00:02 | 00:00:16 |
| 2 | Dermatology | 00:02:28 | 07:41:03 | 00:00:00 | 00:00:05 | 00:00:04 | 00:00:06 |
| 3 | Hepatitis | 00:00:01 | 00:01:52 | 00:00:00 | 00:00:00 | 00:00:00 | 00:00:07 |
| 4 | Ionosphere | 00:12:14 | 39:54:06 | 00:00:02 | 00:00:21 | 00:00:14 | 00:02:16 |
| 5 | Sonar | 01:59:24 | - | 00:00:02 | 00:00:30 | 00:00:09 | 00:04:30 |
| 6 | Spambase | 03:34:44 | - | 00:01:03 | 00:29:40 | 00:03:26 | 02:23:42 |
| 7 | Spectfheart | 00:22:53 | 68:18:37 | 00:00:00 | 00:00:07 | 00:00:03 | 00:00:52 |
| 8 | Wdbc | 00:04:58 | 13:21:14 | 00:00:00 | 00:00:03 | 00:00:02 | 00:00:17 |

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

| **No.** | **Dataset name** | $\boldsymbol{N}$ | | | $\boldsymbol{S}_n$ | | | $\neq C$ | $\neq Pte$ | $\boldsymbol{S}_{2n}$ | | | $\neq C$ | $\neq Pte$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ***#R*#C*** | $\boldsymbol{P}_{tr}$ | $\boldsymbol{P}_{te}$ | ***#R*#C*** | $\boldsymbol{P}_{tr}$ | $\boldsymbol{P}_{te}$ | | | ***#R*#C*** | $\boldsymbol{P}_{tr}$ | $\boldsymbol{P}_{te}$ | | |
| 1 | Bands | 52.20 | 76.17 | 72.80 | 63.60 | 73.63 | 70.60 | -11.40 | 2.20 | 70.62 | 73.21 | 69.67 | -18.42 | 3.13 |
| 2 | Dermato. | 198.05 | 98.03 | 96.07 | 229.72 | 97.48 | **96.08** | -31.66 | **-0.01** | 178.67 | 91.28 | 89.81 | 19.39 | 6.26 |
| 3 | Hepatitis | 26.16 | 95.83 | 88.44 | 25.60 | 96.48 | **88.64** | 0.56 | **-0.20** | 21.16 | 96.53 | **89.67** | 5.00 | **-1.23** |
| 4 | Ionosphere | 90.33 | 95.35 | 90.22 | 108.00 | 94.83 | **90.23** | -17.67 | **-0.01** | 66.75 | 93.89 | **91.46** | 23.58 | **-1.24** |
| 5 | Sonar | 79.76 | 88.39 | 76.80 | 67.98 | 88.99 | **80.41** | 11.78 | **-3.61** | 61.98 | 87.25 | **79.69** | 17.78 | **-2.89** |
| 6 | Spambase | 30.00 | 84.83 | 84.62 | 37.60 | 86.57 | **86.10** | -7.60 | **-1.48** | 26.92 | 86.34 | **85.98** | 3.08 | **-1.36** |
| 7 | Spectfheart | 22.52 | 83.08 | 81.28 | 26.80 | 84.03 | 80.80 | -4.28 | 0.48 | 35.32 | 84.41 | **82.02** | -12.80 | **-0.74** |
| 8 | Wdbc | 37.35 | 97.62 | 96.96 | 55.00 | 97.92 | 96.60 | -17.65 | 0.36 | 34.35 | 97.12 | 95.66 | 3.00 | 1.30 |
| Mean | | 67.05 | 89.91 | 85.90 | 68.31 | **90.00** | **86.20** | | | **61.97** | 88.75 | 85.50 | | |

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

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 07:41:03 or 27,663 seconds, which is greater than 5,532 and 4,610 times in case of the feature's quantities of $\sqrt{n}+1$ (05 seconds) and $\sqrt{2n}+1$ (06 seconds) respectively. The "-" characters 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 sectionfor the FRBCS design are shown in the Table 4, where noting that $\#R$,$\#C$and $\#R*\#C$ are the number of fuzzy rules, the number of conditions and the complexity of the extracted fuzzy rule setrespectively; $P_{tr}$ := the performance in the training phase and $P_{te}$ :=the performance in the testing phase; The $\neq C$ and $\neq Pte$ 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 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).

| VS | R$^+$ | R$^-$ | E. *P*-value | A. *P*-value | Conf. Inte. | Exact. Conf. | Hypothesis |
|---|---|---|---|---|---|---|---|
| $S_{2n}$ | 30 | 6 | 0.10938 | 0.080058 | [-17.9715 , 0.192] | 0.92188 | Not rejected |
| $S_n$ | 10 | 26 | $\geq$ 0.2 | 1 | [-6.71055 , 17.784] | 0.92188 | Not rejected |

Table 5: The comparison result of the fuzzy rule complexitiesusing 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 |
|---|---|---|---|---|---|---|---|
| $S_{2n}$ | 16 | 20 | $\geq$ 0.2 | 1 | [-1.625 , 1] | 0.92188 | Not rejected |
| $S_n$ | 19 | 17 | $\geq$ 0.2 | 0.833635 | [-1.36 , 2.515] | 0.92188 | Not rejected |

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

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 ofthose extracted from the datasets with the subsets of selected features in both cases of the feature's quantities of $\sqrt{n}+1$ and $\sqrt{2n}+1$ 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 [21] 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, it is not needed to take the complexity of the FRBCS into account in the comparisons.

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

It does not make sense when comparing the experimental results of the FRBCS with other classical classification methods based on the statistical and probabilistic approaches because those methods do not have the complexity objectives in the learning processes. To show the efficiency of the FRBCS design based on the hedge algebras methodology, the proposed method in this paper is compared with three well-known learning algorithms regardless of the complexity objectives. The compared methods are Naïve Bayes, a probabilistic classifier based on Bayesian model; nearest neighbour algorithm (k-NN), an example is classified with the majority class of its $k$ nearest neighbours; and support vector machine (SVM) with polynomial kernel [5, 18, 20]. The datasets used to validate are shown in the Table 7 and the experimental results of the classification performances on the test sets are shown in the Table 8.

| **No.** | **Dataset name** | **Number of attributes** | **Number of classes** | **Number of patterns** | **Number of selected features** |
|---|---|---|---|---|---|
| 1 | Dermatology | 34 | 6 | 358 | 10 |
| 2 | Lung cancer | 56 | 3 | 32 | 8 |
| 3 | Prostate cancer | 12,600 | 2 | 102 | 30 |
| 4 | Sonar | 60 | 2 | 208 | 12 |
| 5 | wdbc | 30 | 2 | 569 | 9 |
| 6 | wpbc | 33 | 2 | 198 | 10 |
| 7 | Zoo | 17 | 7 | 101 | 7 |

Table 7: The high dimensional datasets used to compare with other algorithms

| **No.** | **Dataset name** | **Our method** | **SVM** | **Naive Bayes** | **k-NN** |
|---|---|---|---|---|---|
| 1 | Dermatology | 97.21 | 97.01 | **97.82** | 96.45 |
| 2 | Lung cancer | **90.83** | 83.33 | 85.83 | 80.00 |
| 3 | Prostate cancer | 97.09 | **99.00** | 98.09 | 98.09 |
| 4 | Sonar | 83.64 | **85.50** | 69.7 | 84.00 |
| 5 | wdbc | 97.02 | **97.60** | 93.10 | 96.80 |
| 6 | wpbc | **81.34** | 81.20 | 69.4 | 78.80 |
| 7 | Zoo | **98.67** | 97.8 | 94.50 | 90.5 |

Table 8: The comparison of the classification performances with other algorithms

With all null-hypotheses which cannot be rejected by the Wilcoxon's signed-rank test at level $\alpha = 0.05$, shown in the Table 9, it is impossible to find any meaningful differences between the proposed method and the rest of three learning algorithms. This test result proves that the proposed method presents a good accuracy on the test set in comparison with other well-known algorithms.

| **VS** | **$R^+$** | **$R^-$** | **E. P-value** | **A. P-value** | **Conf. Inte.** | **Exact. Conf.** | **Hypothesis** |
|---|---|---|---|---|---|---|---|
| SVM | 14 | 14 | $\geq$ 0.2 | 0.932647 | [-1.86 , 3.85] | 0.95312 | Not rejected |
| Naïve Bayes | 25 | 3 | 0.07812 | 0.051913 | [-0.61 , 11.94] | 0.95312 | Not rejected |
| k-NN | 22 | 6 | $\geq$ 0.2 | 0.150786 | [-0.39 , 6.685] | 0.95312 | Not rejected |

Table 9: The comparison result of the classification performances of the proposed method and three well-known algorithms using the Wilcoxon's signed rank test at level $\alpha = 0.05$

## 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 hedge algebras methodology for the high-dimensional datasets. By utilizing this technique, the extended method for the fuzzy rule based classifier design based on hedge algebras is 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 saves much execution time than the original one, while retaining the equivalent classification performance as well as the equivalent FRBCS complexity. Furthermore, the proposed method is also compared with three other well-known learning algorithms and the results on the accuracy of the test set are comparable with the results obtained by those compared algorithms.

## ACKNOWLEDGMENT

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

## REFERENCES

[1] R. Alcalá, Y. Nojima, F. Herrera, and H. Ishibuchi, “Generating single granularity-based fuzzy classification rules for multiobjective genetic fuzzy rule selection,” in *IEEE International Conference on Fuzzy Systems, 2009. FUZZ-IEEE 2009.* IEEE, 2009, pp. 1718–1723.

[2] ——, “Multiobjective genetic fuzzy rule selection of single granularity-based fuzzy classification rules and its interaction with the lateral tuning of membership functions,” *Soft Computing*, vol. 15, no. 12, pp. 2303–2318, 2011.

[3] J. Alcala-Fdez, R. Alcala, and F. Herrera, “A fuzzy association rule-based classification model for high-dimensional problems with genetic rule selection and lateral tuning,” *IEEE Transactions on Fuzzy Systems*, vol. 19, no. 5, pp. 857–872, 2011.

[4] U. M. Fayyad and K. B. Irani, “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, pp. 1022–1027.

[5] A. González and R. Pérez, “Improving the genetic algorithm of slave,” *Mathware & Soft Computing*, vol. 16, no. 1, pp. 59–70, 2009.

[6] N. C. Ho, “A topological completion of refined hedge algebras and a model of fuzziness of linguistic terms and hedges,” *Fuzzy Sets and Systems*, vol. 158, no. 4, pp. 436–451, 2007.

[7] N. C. Ho, T. D. Khang, H. V. Nam, and N. H. Chau, “Hedge algebras, linguistic-value logic and their application to fuzzy reasoning,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 7, no. 04, pp. 347–361, 1999.

[8] N. C. Ho and N. V. Long, “Fuzziness measure on complete hedge algebras and quantifying semantics of terms in linear hedge algebras,” *Fuzzy Sets and Systems*, vol. 158, no. 4, pp. 452–471, 2007.

[9] N. C. Ho, W. Pedrycz, D. T. Long, and T. T. Son, "A genetic design of linguistic terms for fuzzy rule based classifiers," *International Journal of Approximate Reasoning*, vol. 54, no. 1, pp. 1–21, 2013.

[10] N. C. Ho and W. Wechler, "Hedge algebras: an algebraic approach to structure of sets of linguistic truth values," *Fuzzy sets and systems*, vol. 35, no. 3, pp. 281–293, 1990.

[11] ——, "Extended hedge algebras and their application to fuzzy logic," *Fuzzy sets and systems*, vol. 52, no. 3, pp. 259–281, 1992.

[12] H. Ishibuchi and T. Yamamoto, "Fuzzy rule selection by multi-objective genetic local search algorithms and rule evaluation measures in data mining," *Fuzzy Sets and Systems*, vol. 141, no. 1, pp. 59–88, 2004.

[13] C. Ji-lin, H. Yuan-long, X. Zong-yi, J. Li-min, and T. Zhong-zhi, "A multi-objective genetic-based method for design fuzzy classification systems," *IJCSNS*, vol. 6, no. 8A, p. 110, 2006.

[14] E. G. Mansoori, M. J. Zolghadri, and S. D. Katebi, "SGERD: A steady-state genetic algorithm for extracting fuzzy classification rules from data," *Fuzzy Systems, IEEE Transactions on*, vol. 16, no. 4, pp. 1061–1071, 2008.

[15] M. K. Pakhira, S. Bandyopadhyay, and U. Maulik, "Validity index for crisp and fuzzy clusters," *Pattern recognition*, vol. 37, no. 3, pp. 487–501, 2004.

[16] ——, "A study of some fuzzy cluster validity indices, genetic clustering and application to pixel classification," *Fuzzy sets and systems*, vol. 155, no. 2, pp. 191–214, 2005.

[17] P. D. Phong, N. C. Ho, and N. T. Thuy, "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 *2013 IEEE RIVF International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF).* IEEE, 2013, pp. 12–17.

[18] J. C. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods - Support Vector Learning.* MIT Press, 1999, pp. 185–210.

[19] M. Salazar Lechuga, "Multi-objective optimisation using sharing in swarm optimisation algorithms," Ph.D. dissertation, University of Birmingham, 2009.

[20] X. Sun, Y. Liu, M. Xu, H. Chen, J. Han, and K. Wang, "Feature selection using dynamic weights for classification," *Knowledge-Based Systems*, vol. 37, pp. 541–549, 2013.

[21] J. H. Zar, *Biostatistical analysis.* Prentice-Hall, Upper Saddle River, NJ, 1999.

*Received on September 25 - 2014*
*Revised on May 21 - 2015*