Rough Set-based Dataset Reduction Method Using Swarm Algorithm and Cluster Validation Function

Kuang-Yu Huang  
Ling Tung University  
kyhuang@teamail.ltu.edu.tw

Ting-Hua Chang  
Unemployed  
tchangipas@gmail.com

Shann-Bin Chang  
Ling Tung University  
changsb@teamail.ltu.edu.tw

Abstract

A Rough Set (RS) based dataset reduction method using SWARM optimization algorithm and a cluster validation function is proposed. In the proposed approach, the user specifies the classification quality required in advance, and the method then finds the attribute reducts and performs attribute discretization to satisfy the desired quality of classification. While many other solutions are possible, the proposed method yields the solution which satisfies the optimal discretization conditions by means of a newly-designed cluster validation index function. The performance of the proposed method is compared with that of two existing attribute reduction and classification methods for eight benchmark datasets. The results confirm that the proposed method provides an effective tool for solving simultaneous attribute reduction and discretization problems.

1. Introduction

Feature Subset Selection (FSS) [1, 2, 3] process refers to choosing a subset of attributes from the set of original attributes in order to simplify a dataset by reducing its dimensionality while retaining the same predictive accuracy. Feature subset selection has a number of important benefits for learning algorithms, including a reduced computation time of the induction algorithm, a lower computational cost and storage requirement, and an improved accuracy of the induced rules. Furthermore, when designing pattern classifiers, feature subset selection is beneficial in improving the overall quality and speed of the classification process.

Rough Set (RS) theory was originally proposed by Pawlak [4] as a mathematical approach to handle imprecision, vagueness and uncertainty inherent in many data analysis problems. Since its introduction, RS theory has been widely used in the development of feature subset selection methods for pattern recognition purposes [5]. In the present study, a RS-based feature subset selection method is used as the basis for a new scheme designed to solve simultaneous attribute reduction and discretization problems.

1.1. RS-based feature subset selection

In RS-based feature subset selection method, the aim is to select a subset from the total set of features (or attributes) from which can predict the decision outputs as well as from the original feature set. This subset of features is called a reduct. The optimal criterion for rough set feature subset selection is to find the smallest or minimal reduct(s) while obtaining high quality classifiers based on the selected features [6]. Three basic approaches to RS-based feature subset selection exist:

(1) Exhaustive search: all reducts can be found from the discernibility functions [7]. The discernibility functions could further be simplified by finding the equivalence relations [8]. However, it has been shown that finding minimal reducts or all reducts are both NP-hard problems [9]. In general, identifying the discernibility among the attributes is an exhaustive solution to the problem, and is only practical for simple datasets. For dataset with tens of attributes and thousands of data, alternative techniques must be considered.

(2) Greedy algorithm: Greedy approaches usually employ the significance of the RS attribute as the guideline for feature subset selection. Such methods use either a forward approach or a backward approach. Forward selection methods start with an empty set or an attribute core, and then add the most significant attributes from the candidate set one at a time until the selected set forms a reduct. By contrast, backward feature subset selection methods start with the full attribute set and remove the attributes one by one while checking each time that the classification accuracy remains unchanged. Besides classification accuracy or quality, several other different significance measures of attributes have been proposed in greedy algorithms [10, 11, 12].

(3) Stochastic methods: Many general purpose stochastic (metaheuristic) methods frameworks have
been developed, e.g., see a recent review [13]. The use of stochastic method(s) in feature subset selection problem has significantly increased the ability of finding very high-quality solutions to hard, practically-relevant combinatorial optimization problems in a reasonable time [7, 14, 15].

1.2. Contribution of this paper

There are two coupled issues in RS-based feature subset selection method: discretization and reducts finding. While many RS-based attribute reduction methods already exist, they only deal with clustered data. For real-valued attributes, an additional step of preprocessing, i.e., data discretization, must be performed before RS classification can take place. Attention [16, 17] has been given to generate a minimal set of decision rules, that is, a general rule induction algorithm is suitable for variable consistency rough set approaches, called VC-DomLEM. By specifying the desired quality and precision of classification, does there exist an optimal way to cluster the data and then select reducts? The way the clusters is determined will change the classification accuracy and attribute dependences, and furthermore will affect the feature subset selection results. What the final reducts would be depends on how the discretization was done. The data discretization and reduct determination processes are so closely related, however, though there are many cluster finding methods, e.g., see [18, 19, 20], none of these methods have been integrated as a part of the attribute reduction algorithms. The lack of linkages between the discretization and feature subset selection tasks arises in part since there is presently no standard metric available for comparing the relative quality of different discretization solutions, and researchers tend to handle discretization and feature subset selection as two independent tasks. Therefore to invent a cluster validation function based on the classification method to find “better” discretization schemes is a premise work.

The present study proposes a RS-based stochastic attribute reduction method, designated as the PFRS-index method, consisting of a Particle Swarm Optimization (PSO) algorithm, the Fuzzy C-Means (FCM) discretization method, RS theory, a Sequential Unconstrained Minimization Technique (SUMT) and a novel cluster validity index function. The proposed method provides the means to identify the optimal discretization scheme and to select appropriate feature subsets consistent with a pre-specified classification quality. A cluster validation index function – MH-index function, modified from Huang index function [21] which focuses on the lower approximation of sets in RS, is utilized for this purpose. While other optimization techniques are possible, in this article SWARM algorithm is used to solve the discretization/attribute reduction problem for its fast converging time. In implementing the SWARM algorithm, the MH-index function is integrated with a penalty function designed to constrain the discretization process in accordance with the pre-specified classification quality. The performance of the proposed PFRS-index method is compared with that of two existing real-valued attribute reduction schemes by performing 10-fold cross validation tests using eight benchmark datasets.

2. Definitions and technical terms

This section presents the various definitions and technical terms used throughout the remainder of the paper.

2.1 Rough set theory

Rough set theory deals with the vagueness and uncertainty inherent in many decision-making processes. In RS theory, data points distinguished by the same information are regarded as indiscernible and this indiscernibility of the data points is handled using the concept of approximation of sets. The upper and lower approximation of sets of \( X \subseteq U \) and \( R \subseteq C \) are denoted as \( \overline{R}(X) \) and \( R(X) \), respectively, and are defined as follows:

\[
\overline{R}(X) = \{ x \in U \mid [x]_R \cap X \neq \emptyset \},
\]

\[
R(X) = \{ x \in U \mid [x]_R \subseteq X \}.
\]

Here, the data points in \( \overline{R}(X) \) are certain members of \( X \), while the data points in \( R(X) \) are possible members of \( X \).

Having determined the lower and upper approximation of sets, the accuracy of the classification results could be quantified as

\[
\alpha_c = \frac{|R(X)|}{|\overline{R}(X)|}
\]

where \(|R(X)|\) and \(|\overline{R}(X)|\) are the cardinalities of the lower and upper approximation of sets, respectively. Quality of classification, which indicates the proportion of objects in the universe \( U \) for which a classification based on decision attribute \( D \) is possible, could be defined as follows:

\[
\gamma(R, D) = \frac{\sum [x]_D : X \in [x]_D}{|U|}.
\]

In the expressions above, the unions are taken for \( x \in [x]_D \). In other words, this process involves combining all the positive regions and summing up the
number of objects involved in such a combination. Quality of classification is used operationally to define and extract reducts, which is the kernel part of RS theory in the application to data mining and rule construction.

2.2 Swarm algorithm

PSO is a computational technique for solving combinatorial-type optimization problems [22]. In PSO, the particle positions represent candidate solutions to the \( n \)-dimensional problem of interest and the particle movements represent the search for a better solution. The position of the \( k \)-th particle is represented by the position vector \( x_k = (x_{k_1}, x_{k_2}, \ldots, x_{k_n}) \), while the velocity of the \( k \)-th particle is represented by the velocity vector \( v_k = (v_{k_1}, v_{k_2}, \ldots, v_{k_n}) \). During the iterative search process, the particles successively adjust their positions in accordance with two poles, namely their personal best position and the global best position. For each particle, the personal best position \( P_i = (P_{i_1}, P_{i_2}, \ldots, P_{i_n}) \) is the previous particle position which returns the highest fitness value thus far. Meanwhile, the global best position \( P_g = (P_{g_1}, P_{g_2}, \ldots, P_{g_n}) \) is the position among all the previous particle positions which returns the highest fitness value thus far.

In the present study, the particle velocity and particle position are updated using the so-called “constrict factor method” proposed by Clerc [23], i.e.,

\[
\begin{align*}
    v_{i_j}(t+1) &= \phi (v_{i_j}(t) + c_1 r_1 (P_{i_j} - x_{i_j}(t)) + c_2 r_2 (P_{g_j} - x_{i_j}(t))) \\
    x_{i_j}(t+1) &= x_{i_j}(t) + v_{i_j}(t+1), \quad j = 1, 2, \ldots, n
\end{align*}
\]

where \( r_1 \) and \( r_2 \) are elements from two uniform random sequences in the range \((0,1)\); \( c_1 (2.05) \) is the individual factor; and \( c_2 (2.05) \) is the societal factor; \( C = c_1 + c_2 \); \( \phi = 2/\sqrt{2 - C - \sqrt{C^2 - 4C}} \).

The positions of the particles in each dimension are clamped to a maximum position \( x_{\text{max}} \), where \( x_{\text{max}} \) is specified in advance by the user. If the adjustment from the velocity causes the particle position in any dimension to exceed \( x_{\text{max}} \), the new position of the particle in that dimension is automatically limited to \( x_{\text{max}} \).

In the problem considered in this study, the particles in the PSO algorithm have the form \((I_{a_1}, I_{a_2}, \ldots, I_{a_n}, N)\), where \( I_{a_i} \) denotes the initial random value assigned to the \( i \)-th condition attribute \( a_i \), and \( N \) denotes the number of clusters per attribute. The value of each element in the attribute part of the vector is limited to the interval \([0,1]\), while \( N \) is bounded in the interval \([2, N_{\text{max}}]\), in which 2 represents the minimal number of clusters per attribute and \( N_{\text{max}} \) represents the maximum permissible number of clusters per attribute. Note that in this study \( N_{\text{max}} \) is specified as 10.

The length of the string corresponds to the number of condition attributes in the dataset plus one. (Note that the last plus-one field is used to carry the number of clusters per attribute.) Furthermore, each element of the string represents the value assigned to the corresponding attribute by a rounding function. For the attribute fields, if the value assigned to a particular attribute is less than 0.5, the attribute is dropped from the dimensionality vector; the attribute is retained if its value is greater than or equal to 0.5. The value of the cluster field is also confined to the interval \([0, 1]\). However, this float number is subsequently converted to an integer between 2 to 10 using a linear mapping\(^1\).

For illustration purposes, assume that record \( x_1 \) in the dataset has three condition attributes and the particle vector has the form \((0.3, 0.55, 0.6, 0.18)\). For this particular particle, only the second and third condition attributes are retained since the value of the first attribute is less than 0.5. Furthermore, the number of clusters of the remaining attributes will be 4. In other words, the PSO algorithm generates a candidate reduct with two condition attributes and gives the number of clusters per attribute.

2.3 Fitness function: MH-index function

The PSO algorithm needs a fitness function to evaluate the best particle. In the present study, the fitness function is provided by the modified Huang (MH)-index function. The MH-index function is a cluster validity function which is designed to minimize the number of clusters per attribute and ensures maximum separation between clusters by maximizing the index value. Furthermore, the MH-index function takes explicit account of the classification quality, and is therefore adequate for our purpose. The MH-index function has the form

\[
MH(C,N,\alpha) = \left(\frac{1}{C} \sum_{i=1}^{C} \frac{E_i}{F_{i,N}} \cdot \frac{D_{N_i}}{D_i}\right)
\]

where \( C \) is the number of clusters per condition attribute, and \( \alpha \), is the accuracy of the RS classification

\(^1\) The formula is \(2 + \text{Round}(N \times 8)\). See Section 3.1 for corresponding discussions.
when evaluated in terms of the \(c\)-th decision class. \(F'_{\mathcal{N}_{d}}\) is obtained by \(F'_{\mathcal{N}_{d}} = \sum_{i=1}^{N} E'_i\), where \(E'_i\) is given by

\[
E'_i = \sum_{j=1}^{n} \mu_{C_i}^r(x_j(d)) \|x_j - z'_i\| / \alpha^s,
\]

in which \(z'_i\) is the multidimensional centroid of the lower approximation of sets associated with the \(c\)-th decision class and is obtained by computing the mean values of the conditional and decision attributes of each record within the corresponding sets. \(\mu_{C_i}(x_j(d))\) is the fuzzy membership value of record \(x_j\) in the \(c\)-th class of the decision attribute \(d\), \(m'\) is the fuzzification parameter\(^2\) and \(n\) is the total number of data points in the dataset. \(E_i\) is a constant for a given dataset in which the instances belong to only one cluster. The value of \(E_i\) is obtained by summing the norms of \(\|x_j - z_i\|\), and is used only to prevent the second term from disappearing. Meanwhile, the value of \(D'_{\mathcal{N}_{d}}\) denotes the maximum separation distance amongst the centroids of all the lower approximation of sets associated with the different decision classes, i.e. \(D'_{\mathcal{N}_{d}} = \max_{i,j} \|x_j - z_i\|\).

Finally, \(E_i\) denotes the maximum separation distance amongst all possible pairs of records in the dataset. For a given dataset, \(E_i\) is a constant and is used to prevent the value of the third term from increasing as the number of attributes within the dataset increases or from decreasing as the number of attributes reduces. (Note that a step-by-step illustration of the process involved in calculating the MH-index function is given in the Appendix.)

**Proposition 1:** Using the MH-index function as the fitness function for the PSO algorithm is effective to find redundant attribute(s) in a dataset.

**Proof:** Suppose that dataset \(D\) contains two highly-correlated condition attributes \([c_1 \ c_2]\), where one of these attributes provides only a small additional contribution to the class information relative to that provided by the other. For simplicity, assume that the numerical contents of \(c_1\) and \(c_2\) are identical. Furthermore, assume that each attribute is clustered to \(N\) clusters. A SWARM particle having the pair vector \([1 \ 0]\) or \([0 \ 1]\) will have larger a MH-index function than a particle with pair \([1 \ 1]\) since the former case has a total of \(C = N\) clusters, while the latter case has \(C=2N\) clusters. All other terms in the index formula such as classification accuracy and cluster separation distance are the same. Particle \([1 \ 1]\) will migrate toward state \([1 \ 0]\) or \([0 \ 1]\) during iterations, and thus the redundant attribute will be removed after some iterations. The same argument could be extended to several redundant attributes case. However, the premise is that one has to run enough SWARM iterations and sufficiently large population to increase the chance of the minimal reduct being expressed by one of the particles.

### 2.4 Constrained function: Sequential Unconstrained Minimization Techniques (SUMT)

As discussed in Section 1.2, in the PFRS-index method proposed in this study, the required classification quality is specified in advance. This requirement adds a constraint to the cluster formation procedure. This problem is handled by Sequential Unconstrained Minimization Techniques (SUMT) [24]. The approach to using SUMT is to create a pseudo-objective function of the form

\[
\Phi(C, N_d, \alpha_r, \gamma(R, D)) = MH(C, N_d, \alpha_r) \times (-1 + r_p \Phi' \gamma(R, D))
\]

where \(MH(C, N_d, \alpha_r)\) is the original MH-index function and \(\Phi' \gamma(R, D)\) is an imposed penalty function whose form depends on the SUMT employed. The second term \(r_p \Phi' \gamma(R, D)\) will decrease dramatically if \(\gamma(R, D)\), the quality of classification defined in Section 2.1, is less than the preset value. The scalar \(r_p\), which is set to 0.01, is a multiplier which determines the magnitude of the penalty. For rest of the paper, this pseudo-objective function is simply referred as the (constrained) MH-index function when used in PSO fitness evaluation.

In the present study, the linear extended penalty function revised by Cassis and Schmit [25] is applied. Here the penalty function takes the form

\[
P' \gamma(R, D) = \sum_{j=1}^{n} \tilde{g}_j \gamma(R, D))
\]

\[
\tilde{g}_j \gamma(R, D)) = \begin{cases} 
1 & \text{if } g_j \gamma(R, D)) \leq \varepsilon \\
\frac{1}{\varepsilon} - \frac{1}{\varepsilon} \frac{g_j \gamma(R, D))}{\varepsilon} & \text{if } g_j \gamma(R, D)) > \varepsilon 
\end{cases}
\]

where

\[
\tilde{g}_j \gamma(R, D)) = 3 \left( \frac{g_j \gamma(R, D))}{\varepsilon} - 1 \right) + 3)
\]

**Footnote:**

\(^2\) \(m'\) is set to 2 in this study. This sets the membership value proportional to the reciprocal of the distance.
In this study, a constraint in cluster formation, \( g_i(\gamma(R,D)) = Acc_{\text{min}} - \gamma(R,D) \), is applied. \( Acc_{\text{min}} \) is the preset quality of classification. The parameter \( \varepsilon \) is a small negative number which marks the transition from the interior penalty to the extended penalty. Hafika and Starnes [26] recommend that \( \varepsilon \) be defined by

\[
\varepsilon = -C(r_p)^a \quad \frac{1}{3} \leq a \leq \frac{1}{2}
\]

where \( C \) and \( a \) are set as 0.15 and 0.5, respectively, in this study.

3. PFRS-index Method

The PFRS-index method proposed in this study extends the applicability of the Huang-index method by classifying the dataset with a controlled degree of uncertainty and by rendering the value of the index function insensitive to the number of condition attributes within the dataset. In implementing the PFRS-index method, the PSO algorithm is used to generate candidate reducts for the dataset and to determine the number of clusters per attribute. The details of the PFRS-index method are presented in the following subsections.

3.1 Initial setup

The PSO algorithm commences by generating an initial population of \( P = 40 \) dimensionality vectors. As described in Section 2.2, the particles in the PSO algorithm have the form \((I_{a_1}, I_{a_2}, ..., I_{a_n}, N)\), where \( I_{a_i} \) and \( N \) are generated randomly to start with. For each particle participated in the PSO process the conditional and decision attributes are all partitioned into \( N \) clusters. The dimensionality vector of one of the initial particles is deliberately specified as \((1.0,1.0, ...,1.0, N = 2)\) in order to ensure that the maximum index value of the candidate reducts obtained by the proposed index function is larger than that obtained when all the condition attributes within the dataset are retained. Then the values of the elements within the dimensionality vectors are assigned a value of 0 or 1 using the rounding function \( I_{a_i} = \text{round}(I_{a_i}) \) to represent the initial reduct candidates.

3.2 Fuzzification and RS classification

For each attribute of the reduct candidates, the original continuous values are then discretized using the FCM discretization method [27]. Each fuzzified attribute of the reduct candidates generated by the PSO algorithm is then assigned to the cluster according to the maximum membership value obtained from FCM. Having mapped all the attribute values of all the reducts in the dataset to the corresponding condition or decision attribute clusters, lower and upper approximation of sets associated with each decision class of the reducts are extracted. The accuracy of the RS classification associated with each decision class is then determined by calculating the cardinality ratio of the corresponding lower approximation of sets to the upper approximation of sets. Finally, the centroids of each attribute of the equivalence classes associated with each decision class are calculated by computing the mean attribute values (both condition and decision) of all of the data within the corresponding equivalence classes.

3.3 SWARM and fitness evaluation

Having determined the number of clusters per attribute, the membership function values of all the attributes of all the records, the accuracy of the RS classification, and the centroids of attribute values of the equivalence classes in each decision class, the optimality of the classification solution of each particle is evaluated using the (constrained) MH-index function. In the PFRS-index method, each particle represents one possible dimensionality vector for the condition attributes, and the aim of the (constrained) MH-index function is to identify the particular reduct which optimizes the number of clusters within the dataset, maximizes the separation distance between these clusters, and optimizes the accuracy of approximation.

Having calculated the value of the objective function shown in Section 2.4 for each of the 40 particles in the current population, the particle which yields the maximum MH-index value and satisfies the preset classification quality is identified as the “best” particle. The initial population of candidate reducts is then evolved using the velocity and position updating equations given in Section 2.2. Having modified the particle vectors, the rounding function is applied to update the corresponding dimensionality vectors. New reduct candidates are then generated for the next iteration, and the fitness of each of the new reduct candidates is once again evaluated using the constrained MH-index function. The SWARM evolution process continues iteratively in this way until the termination criterion is specified, i.e., \( N\text{Loop} = 50 \) loops. Once the PSO termination criterion has been satisfied, all dimensionality vectors generated by the PSO algorithm are examined in order to identify the

\[3 \text{ See the Discussion section for the determination of this particular choice of loops as the termination criterion.}\]
vector which yields the maximum value of the MH-index function. The attribute subset specified by this particular dimensionality vector is then taken as the optimal reduct of the dataset, and the corresponding value of N is taken as the optimal number of clusters per attribute.

4. Benchmark datasets study

The performance of the proposed PFRS-index method was evaluated using eight different datasets. Seven of the datasets were extracted from the UCI machine learning repository [28] (i.e., Breast Tissue, Ecoli, Glass, Libras, Segment, Bench, and Vowel), while one dataset (Vehicle) was taken from the Statlog repository [29]. The number of data entries and condition attributes in each dataset are summarized in Table 1.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Data Entries</th>
<th>Condition attributes</th>
<th>Classes numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast Tissue</td>
<td>106</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>Ecoli</td>
<td>336</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Glass</td>
<td>214</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>Libras</td>
<td>360</td>
<td>90</td>
<td>15</td>
</tr>
<tr>
<td>Segment</td>
<td>2310</td>
<td>16</td>
<td>7</td>
</tr>
<tr>
<td>Bench</td>
<td>208</td>
<td>60</td>
<td>2</td>
</tr>
<tr>
<td>Vowel</td>
<td>528</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>Vehicle</td>
<td>846</td>
<td>18</td>
<td>4</td>
</tr>
</tbody>
</table>

4.1 Results

For each dataset, feature subset selection was performed and the classification quality of the remaining subset of attributes was tested by means of a 10-fold cross validation process. The results obtained from the PFRS-index method for each dataset are presented in Table 2. In general, the results confirm the ability of the PFRS-index method to identify an attribute subset which retains the original class information.

The entries in the preset quality requirement column of Table 2 specify the minimal acceptable quality of classification and are taken from the classification results obtained using other methods (see Table 3). The value is deliberately chosen from the inferior quality between CART-DT or CART-SVM in order to demonstrate that even with a more stringent quality of classification requirement, the obtained quality of classification is still better than that of other methods. It should be noted that the entries within the other columns in Table 2 (i.e., the Average Number of Clusters, the Average Number of Remaining Attributes, and the Average Quality of Classification) represent the average 10-fold results.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>PQR</th>
<th>ANC</th>
<th>ANRA</th>
<th>AQC</th>
<th>MR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast Tissue</td>
<td>0.86</td>
<td>6.9</td>
<td>6.3</td>
<td>0.973</td>
<td>2.7%</td>
</tr>
<tr>
<td>Ecoli</td>
<td>0.76</td>
<td>5.5</td>
<td>7</td>
<td>0.994</td>
<td>0.6%</td>
</tr>
<tr>
<td>Glass</td>
<td>0.75</td>
<td>5.2</td>
<td>7.6</td>
<td>0.981</td>
<td>1.9%</td>
</tr>
<tr>
<td>Libras</td>
<td>0.36</td>
<td>2</td>
<td>64.7</td>
<td>0.972</td>
<td>2.8%</td>
</tr>
<tr>
<td>Segment</td>
<td>0.98</td>
<td>4.7</td>
<td>13.4</td>
<td>0.983</td>
<td>1.7%</td>
</tr>
<tr>
<td>Bench</td>
<td>0.69</td>
<td>2</td>
<td>30.3</td>
<td>1.000</td>
<td>0.0%</td>
</tr>
<tr>
<td>Vowel</td>
<td>0.89</td>
<td>4</td>
<td>9.2</td>
<td>0.991</td>
<td>0.9%</td>
</tr>
<tr>
<td>Vehicle</td>
<td>0.73</td>
<td>3.1</td>
<td>17.7</td>
<td>0.981</td>
<td>1.9%</td>
</tr>
</tbody>
</table>

1. PQR: Preset Quality Requirement.
2. ANC: Average Number of Clusters.
3. ANRA: Average Number of Remaining Attributes.
4. AQC: Average Quality of Classification.
5. MR: Misclassification Rate

4.2 Comparison of PFRS-index method and other methods

The performance of the PFRS-index method was compared with that of two other feature subset selection and classification methods, namely CART-DT and CART-SVM [30, 31, 32]. In every case, the average “quality of classification” was determined by means of a ten-fold cross validation test. Note that the term “quality of classification” is used here to represent the proportion of objects which are accurately classified in each of the three methods. (Note that in the case of the PFRS-index method, the quality of classification is indicated by the proportion of objects contained within the lower approximation of sets.) In implementing the two methods used for comparison purposes, feature subset selection was performed using the CART decision tree algorithm [30], and the training datasets were then classified based on the remaining attributes using the Decision Tree (DT) and Support Vector Machine (SVM) classifiers, respectively. Finally, the trained classifiers were applied to the test datasets and the corresponding quality of classification computed. The classification results for the three schemes are summarized in Table 3.
of clusters per attribute, and no effect on the classification processes. The method provides more attributes case, and the result shows no predictive power and attributes. The result showed using the Ecoli dataset. In performing the tests, the appropriate to add these green words) was investigated performance of the PFRS-index function method (Is it the feature subset selection and classification by the (1/N) term. implies rejection of such trivial discretization scheme classification. The form of MH-index function also will fail the cross validation test on quality of reduct by definition, but has as the consequence only the particles with satisfactory quality of classification would become the candidate of the best particle. On the other hand, the preset qualities appear to have no effect on the performance of attribute reduction.

5. Discussion

A randomly-selected attribute column in the Breast Tissue dataset was duplicated and added to the dataset in order to test the ability of the PFRS-index method to remove the redundant attributes. The result showed that the PFRS-index method can indeed remove this redundant attribute during SWARM process. This examination for redundancy removal was extended to several duplicated attributes case, and the result showed the PFRS-index method is sufficient to remove most of the redundant attributes. These results are in line with Proposition 1 and confirm the suitability of the parameter settings of Population=40 and Nloop=50 assigned in the SWARM optimization procedure. Given the high quality of classification, the remaining attributes could be considered as a psudo-reduct of the original dataset.

The PFRS-index method also prevents the case where all but one of the attributes is eliminated and the remaining attribute is then clustered on an entry-by-entry basis. In such a case, the only left attribute is a reduct by definition, but has no predictive power and will fail the cross validation test on quality of classification. The form of MH-index function also implies rejection of such trivial discretization scheme by the (1/N) term.

The effect of the preset quality of classification on the feature subset selection and classification performance of the PFRS-index function method (Is it appropriate to add these green words) was investigated using the Ecoli dataset. In performing the tests, the preset classification quality was assigned values of 0.9, 0.8, 0.7, 0.6 and 0.5 in the 10-fold cross-validation runs to see how the initial setting could affect the final quality of classification, respectively, and 90% of the original data entries were used in each run. The corresponding results are shown in Table 4.

### Table 3 Qualities/Standard deviations of classification of test datasets obtained from three different methods.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>PFRS Preset Quality</th>
<th>PFRS-index method</th>
<th>CART - DT</th>
<th>CART - SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast Tissue</td>
<td>0.86</td>
<td>0.973/8.62%</td>
<td>0.670/16.58%</td>
<td>0.139/9.69%</td>
</tr>
<tr>
<td>Ecoli</td>
<td>0.76</td>
<td>0.994/1.86%</td>
<td>0.702/6.69%</td>
<td>0.762/9.13%</td>
</tr>
<tr>
<td>Glass</td>
<td>0.75</td>
<td>0.981/7.02%</td>
<td>0.645/8.83%</td>
<td>0.658/9.38%</td>
</tr>
<tr>
<td>Libras</td>
<td>0.36</td>
<td>0.972/3.93%</td>
<td>0.494/8.57%</td>
<td>0.183/6.03%</td>
</tr>
<tr>
<td>Segment</td>
<td>0.90</td>
<td>0.968/2.29%</td>
<td>0.936/1.44%</td>
<td>0.597/3.24%</td>
</tr>
<tr>
<td>Bench</td>
<td>0.69</td>
<td>1.000/0.00%</td>
<td>0.712/7.95%</td>
<td>0.625/10.27%</td>
</tr>
<tr>
<td>Vowel</td>
<td>0.89</td>
<td>0.991/2.04%</td>
<td>0.716/6.31%</td>
<td>0.884/4.84%</td>
</tr>
<tr>
<td>Vehicle</td>
<td>0.73</td>
<td>0.988/1.67%</td>
<td>0.706/5.89%</td>
<td>0.704/5.48%</td>
</tr>
</tbody>
</table>

5. Conclusion

This study has presented a method designated as the PFRS-index method for solving joint feature subset selection and classification problems. The scheme comprises a SWARM optimization algorithm and an MH-index function (used as the fitness function in the optimization procedure). In performing the optimization procedure, the goal is to determine the reduct and number of clusters per attribute which maximize the value of the MH-index function. In contrast to existing feature subset selection and classification methods, the PFRS-index method proposed in this study provides the means to specify the desired quality of classification (i.e., the proportion of data being accurately classified) in advance of the feature subset selection and classification processes while other methods of feature subset selection do not provide this option. Thus this method provides more flexibility and cleverness than other methods.

### Table 4 Effect of preset quality of classification on feature subset selection performance of PFRS-index method. (Note that results correspond to Ecoli dataset.)

<table>
<thead>
<tr>
<th>Preset Quality</th>
<th>Average Number of Clusters of 10 Runs</th>
<th>Average Number of Remaining Attributes over 10 Runs</th>
<th>Average Quality of Classification over 10 Runs</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9</td>
<td>6.2</td>
<td>7</td>
<td>0.971</td>
</tr>
<tr>
<td>0.8</td>
<td>5.7</td>
<td>7</td>
<td>0.945</td>
</tr>
<tr>
<td>0.7</td>
<td>5.5</td>
<td>7</td>
<td>0.934</td>
</tr>
<tr>
<td>0.6</td>
<td>5.2</td>
<td>7</td>
<td>0.904</td>
</tr>
<tr>
<td>0.5</td>
<td>5.1</td>
<td>7</td>
<td>0.894</td>
</tr>
</tbody>
</table>

It appears that a higher preset quality would result in larger average number of clusters per attribute, and consequently results in a better average quality of classification. This trend is expected since the penalty function is added to the MH-index function as the fitness evaluation; any violation to the preset quality constraint would reduce the fitness value dramatically, and as the consequence only the particles with satisfactory quality of classification would become the candidate of the best particle. On the other hand, the preset qualities appear to have no effect on the performance of attribute reduction.
Eight datasets are studied by three methods, including the proposed one, to exam the over all classification ability after feature subset selection. For the methods under comparison, the CART-DT is a wrapper method which combines the feature subset selection and classifier, and for CART-SVM the feature subset selection and classifier utilizes two different techniques. In general, the results have confirmed that the PFRS-index method provides a better classification performance following feature subset selection than either of the two CART-based methods.

As the future prospect, this research could be extended to compare the results of quality of classification obtained using the proposed method with those using another rough set based approach, VC-DomLEM.

7. Acknowledgements

This study was financially supported by the Research Grant MOST 103-2410-H-275 -004 - from Taiwan’s Ministry of Science and Technology.

8. References

Appendix: Step-by-step illustration of MH-index function derivation process

This section illustrates the derivation of the MH-index function value for a simple hypothetical dataset comprising just four entries. Each entry is assumed to have two condition attributes, \( a_1 \), \( a_2 \), and one decision attribute, \( d \), i.e., \( x_i(a_1, a_2, d) \). Furthermore, it is assumed that each condition attribute can be partitioned into 2 clusters and the label information of the decision attribute is pre-determined. Let the four entries be defined as \( x_1(0.90, 2.30, 0.2) \), \( x_2(1.10, 2.20, 2) \), \( x_3(1.45, 2.45, 1) \), and \( x_4(1.55, 2.55, 1) \), respectively. In calculating the value of the MH-index function, the real-valued condition attributes in the hypothetical dataset are initially discretized using FCM. The membership function values of each attribute of each entry are summarized in Table A(a). For each entry, the appropriate condition attribute clusters are obtained in accordance with the maximum membership function values (shown in Table A(b)). As shown, the discretized vectors of the four objects are \( x_1(2, 2, 2) \), \( x_2(2, 2, 2) \), \( x_3(1, 1, 1) \), and \( x_4(1, 1, 1) \), respectively.

Table A(a) Membership function values of each attribute of each entry

<table>
<thead>
<tr>
<th>Index of entries</th>
<th>Condition attributes</th>
<th>Decision Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a_1 )</td>
<td>( a_2 )</td>
<td>( d )</td>
</tr>
<tr>
<td>1</td>
<td>0.025</td>
<td>0.975</td>
</tr>
<tr>
<td>2</td>
<td>0.063</td>
<td>0.937</td>
</tr>
<tr>
<td>3</td>
<td>0.988</td>
<td>0.012</td>
</tr>
<tr>
<td>4</td>
<td>0.992</td>
<td>0.008</td>
</tr>
</tbody>
</table>

Having determined the appropriate condition attribute clusters, the RS classification process is applied; yielding the upper and lower approximation of sets shown in Table A(b). The accuracy of approximation associated with each decision class is then obtained by determining the cardinality ratio of the corresponding lower approximate sets to the upper approximation sets. In the present example, the accuracies of approximation for the two decision classes are therefore \( \alpha_1 = 2/2 = 1.000 \) and \( \alpha_2 = 2/2 = 1.000 \), respectively.

Then, the multi-dimensional centroids of each equivalence class in the lower approximation of sets associated with each decision class are determined by calculating the mean attribute values (condition attributes only) of all the objects within the corresponding equivalence class. Thus, the centroids of the equivalence classes associated with the two decision classes are obtained as

\[
\overline{x}_1 = \frac{\text{mean}(x)}{\text{mean}(x_i, x_j)} = \frac{(0.90 + 1.10)/2, (2.30 + 2.20)/2} = (1.00, 2.25) \quad \text{and} \quad \overline{x}_2 = \frac{\text{mean}(x)}{\text{mean}(x_i, x_j)} = (1.50, 2.50),
\]

respectively.

Table A(b). Lower approximation of sets and upper approximation of sets associated with \( c \)-th decision attribute.

<table>
<thead>
<tr>
<th>Index of entries</th>
<th>Lower approximation of sets</th>
<th>Upper approximation of sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( \mathcal{R}(x : C_a(x) = c, x \in X) )</td>
<td>( \mathcal{R}(x : C_a(x) = 2, x \in X) )</td>
</tr>
<tr>
<td>2</td>
<td>( \mathcal{R}(x : C_a(x) = 2, x \in X) )</td>
<td>( \mathcal{R}(x : C_a(x) = 1, x \in X) )</td>
</tr>
<tr>
<td>3</td>
<td>( \mathcal{R}(x : C_a(x) = 1, x \in X) )</td>
<td>( \mathcal{R}(x : C_a(x) = c, x \in X) )</td>
</tr>
</tbody>
</table>

The optimality of the classification results is then evaluated using the MH-index function (i.e.,

\[
\text{MH}(C, N_d, a_i) = \frac{1}{C} \times \frac{\sum_{i=1}^{N_d} \mu_{i}^m}{\sum_{i=1}^{N_d} D_{i}}
\]

). Note that in illustrating the derivation of \( F_{N_{ij}} \) (where \( F_{N_{ij}} = \sum_{i=1}^{N_d} E_i \)) and the following discussions arbitrarily consider the computation of \( E_i \). The first object in the dataset, \( x_i \), has condition attribute values of \( x_i(0.90, 2.30) \). As a result, \( (x_i(a_1) - \overline{x}_1(a_1)) = (0.90 - 1.50) = -0.60 \) and \( (x_i(a_2) - \overline{x}_2(a_2)) = (2.30 - 2.50) = -0.20 \). Thus, the vector \( x_i = x_i - \overline{x}_i \) has the form \( [x_{11}(a_1), x_{12}(a_2)]\), \(-0.60, -0.20\), and the corresponding norm is equal to \( \| x_i - \overline{x}_i \| = \sqrt{(-0.60)^2 + (-0.20)^2} = 0.632 \). Let the fuzzification parameter \( m' \) be assigned a value of 2.0. Applying the notation \( \| x_i - \overline{x}_i \| = \mu_{i}^{m'}(x_i(d)) \times \| x_i - \overline{x}_i \| \), the effect of object \( x_i \) on \( \overline{x}_i \), i.e., \( |\mu_{i}^{m'}| \), is obtained by

\[\text{Step-by-step illustration of MH-index function derivation process}\]

This section illustrates the derivation of the MH-index function value for a simple hypothetical dataset comprising just four entries. Each entry is assumed to have two condition attributes, \( a_1, a_2 \), and one decision attribute, \( d \), i.e., \( x_i(a_1, a_2, d) \). Furthermore, it is assumed that each condition attribute can be partitioned into 2 clusters and the label information of the decision attribute is pre-determined. Let the four entries be defined as \( x_1(0.90, 2.30, 0.2) \), \( x_2(1.10, 2.20, 2) \), \( x_3(1.45, 2.45, 1) \), and \( x_4(1.55, 2.55, 1) \), respectively. In calculating the value of the MH-index function, the real-valued condition attributes in the hypothetical dataset are initially discretized using FCM. The membership function values of each attribute of each entry are summarized in Table A(a). For each entry, the appropriate condition attribute clusters are obtained in accordance with the maximum membership function values (shown in Table A(b)). As shown, the discretized vectors of the four objects are \( x_1(2, 2, 2) \), \( x_2(2, 2, 2) \), \( x_3(1, 1, 1) \), and \( x_4(1, 1, 1) \), respectively.

Table A(a) Membership function values of each attribute of each entry

<table>
<thead>
<tr>
<th>Index of entries</th>
<th>Condition attributes</th>
<th>Decision Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a_1 )</td>
<td>( a_2 )</td>
<td>( d )</td>
</tr>
<tr>
<td>1</td>
<td>0.025</td>
<td>0.975</td>
</tr>
<tr>
<td>2</td>
<td>0.063</td>
<td>0.937</td>
</tr>
<tr>
<td>3</td>
<td>0.988</td>
<td>0.012</td>
</tr>
<tr>
<td>4</td>
<td>0.992</td>
<td>0.008</td>
</tr>
</tbody>
</table>

The optimality of the classification results is then evaluated using the MH-index function (i.e.,

\[
\text{MH}(C, N_d, a_i) = \frac{1}{C} \times \frac{\sum_{i=1}^{N_d} \mu_{i}^m}{\sum_{i=1}^{N_d} D_{i}}
\]

). Note that in illustrating the derivation of \( F_{N_{ij}} \) (where \( F_{N_{ij}} = \sum_{i=1}^{N_d} E_i \)) and the following discussions arbitrarily consider the computation of \( E_i \). The first object in the dataset, \( x_i \), has condition attribute values of \( x_i(0.90, 2.30) \). As a result, \( (x_i(a_1) - \overline{x}_1(a_1)) = (0.90 - 1.50) = -0.60 \) and \( (x_i(a_2) - \overline{x}_2(a_2)) = (2.30 - 2.50) = -0.20 \). Thus, the vector \( x_i = x_i - \overline{x}_i \) has the form \( [x_{11}(a_1), x_{12}(a_2)]\), \(-0.60, -0.20\), and the corresponding norm is equal to \( \| x_i - \overline{x}_i \| = \sqrt{(-0.60)^2 + (-0.20)^2} = 0.632 \). Let the fuzzification parameter \( m' \) be assigned a value of 2.0. Applying the notation \( \| x_i - \overline{x}_i \| = \mu_{i}^{m'}(x_i(d)) \times \| x_i - \overline{x}_i \| \), the effect of object \( x_i \) on \( \overline{x}_i \), i.e., \( |\mu_{i}^{m'}| \), is obtained by.

1491
multiplying $\|x_i - z_i\|$ by the square of the corresponding membership function value, i.e., $\mu_i^2(x_j(a)) = 0.000^2 = 0.000$. Thus, $\|x_i\|$ has a value of 0.000. $\|z_1\|, \|z_2\|$, and $\|z_3\|$ are calculated using the same procedure. The corresponding results are shown in Table A(c). The value of $E_1$ is thus obtained as $E_1 = \frac{\sum_{j=1}^{4} \mu_i^2(x_j(d)) \|x_j - z_i\|}{\sum_{j=1}^{4} \mu_i^2(x_j(d))} = \frac{\|x_1\| + \|x_2\| + \|x_3\|}{\|x_1\| + \|x_2\| + \|x_3\|} = (0.000 + 0.000 + 0.071 + 0.071)/1.000 = 0.142$. Utilizing the same approach, the value of $E_2$ is obtained as 0.224. $F'_{i,j}$ is thus found to have a value of $F'_2 = \sum_{c=1}^{2} E'_c = 0.366$.

The centroid of the candidate reduct for the illustrative dataset is given by $z_1 = mean(x_i \mid x \in \{x_i, i = 1, 2, ..., 4\})$. As a result, the centroid $z_1$ (calculated by the arithmetic mean function $mean(x_i \mid x \in \{x_i, i = 1, 2, ..., 4\})$) has attribute values of $0.90$ and $1.10$ for the condition attributes within the dataset reduces. The membership function value, i.e., $(x_1(a_1) - x_1(a_2)) = (0.90 - 1.250) = -0.350$ and $(x_1(a_2) - z_1(a_2)) = (2.30 - 2.375) = -0.075$. Therefore, the vector $x_1 = x_1 - z_1$ has the form $[x_1(a_1), x_1(a_2)] = [0.350, -0.075]$, and the corresponding norm is equal to

$$\|x_1 - z_1\| = \sqrt{x_1(a_1)^2 + x_1(a_2)^2} = \sqrt{(-0.350)^2 + (-0.075)^2} = 0.358.$$ 

Similarly, the norms of $\|x_2 - z_2\|$, $\|x_3 - z_3\|$ and $\|x_4 - z_4\|$ are found to be 0.230, 0.214 and 0.347, respectively. The value of $E_i$ in the MH-index function is then obtained by summing the norms of $\|x_j - z_i\|$ where $j = 1, 2, ..., 4$, yielding a value of $E_i = 1.149$.

The value of $D'_d$, in the MH-index function is obtained by calculating the maximum separation distance between the centroids of the equivalence classes in the lower approximation of sets associated with the first and second decision classes. As shown above, these centroids are given by $z_1(1.50, 2.50)$ and $z_2(1.00, 2.25)$, respectively. Thus, the vector $z_{12} = z_1 - z_2$ which maximizes the value of $D'_d = \max_{i,j} \|E_{ij} - z_i\|$ has the form $[z_{12}(a_1), z_{12}(a_2)] = [0.50, 0.25]$, and the corresponding norm is $\sqrt{0.50^2 + 0.25^2} = 0.559$. Factor $D'_d$, in the MH-index function is a constant term and is used only to prevent the value of the third term, $D'_d$, from decreasing as the number of condition attributes within the dataset reduces. The value (norm) of $D'_d$ for this illustrative dataset is equal to

$$\|x_1 - x_4\| = \|x_1(a_1) - x_4(a_1)\|^2 + (x_1(a_2) - x_4(a_2))^2 = \sqrt{(-0.650)^2 + (-0.250)^2} = 0.696.$$ 

Finally, the MH-index function ($MH(C, N_d, \alpha_d) = \frac{1}{C} \sum_{i=1}^{C} \frac{E_i}{D'_d}$) is found to have a value of 1.264, where $C = N_d = 2$, $E_i = 1.149$, $F' = 0.366$, $D'_d = 0.559$ and $D'_d = 0.696$.

<table>
<thead>
<tr>
<th>$\bar{z}_j$</th>
<th>$\bar{z}_c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.000</td>
</tr>
<tr>
<td>2</td>
<td>0.000</td>
</tr>
<tr>
<td>3</td>
<td>0.071</td>
</tr>
<tr>
<td>4</td>
<td>0.071</td>
</tr>
</tbody>
</table>

Table A(c) Values of $\|x'_j\| = \mu_i^2(x_j(d)) \times \|x_j - z'_i\|$.