Research Article

A Novel Strategy for Minimum Attribute Reduction Based on Rough Set Theory and Fish Swarm Algorithm

Yuebin Su¹,² and Jin Guo¹

¹School of Information Science and Technology, Southwest Jiao Tong University, Chengdu 610031, China
²School of Mathematics and Statistics, Sichuan University of Science & Engineering, Zigong 643000, China

Correspondence should be addressed to Yuebin Su; mathsyb@163.com

Received 28 March 2017; Revised 25 June 2017; Accepted 5 July 2017; Published 15 August 2017

Academic Editor: Naveed Ejaz

Copyright © 2017 Yuebin Su and Jin Guo. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

For data mining, reducing the unnecessary redundant attributes which was known as attribute reduction (AR), in particular, reducts with minimal cardinality, is an important preprocessing step. In the paper, by a coding method of combination subset of attributes set, a novel search strategy for minimal attribute reduction based on rough set theory (RST) and fish swarm algorithm (FSA) is proposed. The method identifies the core attributes by discernibility matrix firstly and all the subsets of noncore attribute sets with the same cardinality were encoded into integers as the individuals of FSA. Then, the evolutionary direction of the individual is limited to a certain extent by the coding method. The fitness function of an individual is defined based on the attribute dependency of RST, and FSA was used to find the optimal set of reducts. In each loop, if the maximum attribute dependency and the attribute dependency of condition attribute set are equal, then the algorithm terminates, otherwise adding a single attribute to the next loop. Some well-known datasets from UCI were selected to verify this method. The experimental results show that the proposed method searches the minimal attribute reduction set effectively and it has the excellent global search ability.

1. Introduction

Data mining, which was known as knowledge discovery in database, includes extracting knowledge, discovering new patterns, and predicting the future trends from the amounts of data. Nowadays, with an increasing number of applications in different fields, massive volumes of very high-dimensional data were produced; the data mining faces the great challenge. As known to all, much of datasets contain unnecessary redundant attributes, which not only occupy extensive computing resources but also seriously impact the decision-making process. Reducing the unnecessary redundant attributes becomes very necessary for data mining [1]. Attribute reduction (AR) in the rough set theory (RST) removes redundant or insignificant knowledge with keeping the classification ability of the information system the same as before. It was proposed by Pawlak and Sowinski [2]. Now, RST is widely used in many fields such as machine learning, data mining, and knowledge discovery [3–6].

AR is one of the core problems in RST. In particular, minimal reduction problem is an important part of AR in RST, in which the cardinality of attribute subset is the smallest among all possible reductions. It has been paid much attention by many researchers. One basic solution to find the minimal reducts is to construct a discernibility function and simplify it from the dataset by discernibility matrix [7–9]. Unfortunately, it has been shown that the problem of minimal reduct generation is NP-hard and the run time of generating all reducts is exponential [10]. Recently, because many kinds of NP-hard problems can be solved by heuristic algorithms with increasing computational cost, heuristic attribute reduction algorithm is the main research direction in the field of AR [11].

In general, swarm intelligence algorithm is one kind of heuristic approaches which were used widely for solving attribute reduction problem, including genetic algorithm (GA) [12–14], particle swarm optimization (PSO) [15–18], ant colony optimization (CO) [19, 20], and fish swarm algorithm (FSA) [11, 21, 22]. FSA is a kind of evolutionary algorithm
which was inspired by the natural schooling behaviors of fish to generate candidate solutions for optimization problems, such as random, swarming, following, and preying behaviors. It has a strong ability to avoid local minima in order to achieve a global optimization [23]. Due to its abilities to perform, FSA has received much attention in recent years.

In this paper, a new coding method about the subset of attribute set is proposed. By the coding method, a novel strategy for minimal attribute reduction algorithm based on FSA and RST is proposed. It firstly identifies the core attributes by discernibility matrix. Based on the core attributes, all subsets without containing the core attribute are encoded into an integer by the proposed coding method and an initial population is generated for FSA used to find the optimal set of reducts. The fitness function of a subset is encoded into an integer by the proposed coding method and the maximum fitness value on the bulletin board. In FSA, let $N$ be the population size, the Artificial Fishes (AF) are generated by random function which is represented by a $d$-dimensional position $X_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$, and $X_{i\text{next}}$ is the updated position for iteration $i$. Food satisfaction of $X_i$ is represented as fitness function value $Y_f = F(X_i)$. The Euclidean distance $d_{ij} = \|X_i - X_j\|$ is the distance between $X_i$ and $X_j$. Other parameters include step (representing maximum step length), visual (the visual distance of fish), rand (a random number in $[0,1]$) and $\sigma$ (a crowd factor).

Preying behavior is a basic behavior of FSA. As shown in (4), for $X_j$, we randomly select a random $X_i$ within the current visual scope. If $Y_i < Y_j$, then move a step from $X_i$ to $X_j$. Otherwise, $X_i$ move a step to another random $X_j$ that $Y_i < Y_j$. After a number of trials, if the random $X_j$ that meets $Y_i < Y_j$ is not satisfied, $X_i$ will be replaced with a random position within the visual scope directly. It makes the FSA escape from the local optimal solution. Define the $\text{prey}(X_i)$ function as (4).

Swarming behavior is described as (5). It shows the attraction of the swarm center to the individual. Let $n_f$ be the number of AFs within the current visual scope of $X_i$, and $X^c_i$ is the center position of those neighbors. For the swarm center $X^c_i$, if the food satisfaction is greater and not too crowded (i.e., $\sigma Y_i < Y_i^{\text{max}}/n_f$), then move a step from $X_i$ to $X^c_i$. Otherwise, preying behavior is to identify a next position for the current.

Following behavior is described as (6). Let $X_i^{\text{max}}$ be a AF with the greatest food consistence among AFs in the current visual scope. If the food satisfaction is greater and not too crowded (i.e., $\sigma Y_i < Y_i^{\text{max}}/n_f$), then move a step from $X_i$ to $X_i^{\text{max}}$. Otherwise, preying behavior is to identify a next position for the current.

### 2. Background

#### 2.1. Base Notions of Rough Set Theory

In this section, some basic notions and its proposition will be reviewed in the theory of rough set.

A decision table can be represented as $S = \{U, A, V, f\}$, where $U = \{x_1, x_2, \ldots, x_n\}$ is a nonempty finite set of objects, $A = C \cup D$, where $C$ is a set of condition attribute and $D$ is a decision attribute set, $V$ is the domains of attributes belonging to $A$, and $f : U \times A \rightarrow V$ is a function assigning attribute values to objects in $U$.

For any $R \subseteq C \cup D$, there is an associated indiscernibility relation $\text{IND}(R)$:

$$\text{IND}(R) = \{(x, y) \mid f(x, a) = f(y, a), \forall a \in P, x, y \in U\}. \quad (1)$$

Let $X \subseteq U$; the $R$-lower approximation of $X$ is defined as $\text{RX} = \{x \in U \mid [x]_R \subseteq X\}$ where $[x]_R$ denotes an equivalence class of $\text{IND}(R)$ determined by object $x$. The notation $\text{POS}_R(D)$ refers to the $R$-positive region determined by object $x$. The notation $\text{POS}_R(D) = \bigcup_{x \in U/\text{IND}(D)} \text{RX}$. The $R$-approximation quality with respect to decisions attribute set $D$ is defined as follows:

$$\gamma_R = \frac{\text{POS}_R(D)}{|U|} \quad (2)$$

and the core attribute set is defined as

$$\text{Core} = \{a : \gamma_R[\{a\}] \neq \gamma_R\}. \quad (3)$$
In addition, step and visual parameters play an important role in FSA. They determine the convergence speed of FSA, ∀ 𝑖(1 ≤ 𝑖 ≤ 𝑚)

### Table 1: C(7, 4).

<table>
<thead>
<tr>
<th>n</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(1 2 3 4)</td>
<td>8</td>
<td>15</td>
<td>22</td>
<td>29</td>
<td>36</td>
<td>43</td>
</tr>
<tr>
<td>2</td>
<td>(1 2 3 5)</td>
<td>9</td>
<td>16</td>
<td>23</td>
<td>30</td>
<td>37</td>
<td>44</td>
</tr>
<tr>
<td>3</td>
<td>(1 2 3 6)</td>
<td>10</td>
<td>17</td>
<td>24</td>
<td>31</td>
<td>38</td>
<td>45</td>
</tr>
<tr>
<td>4</td>
<td>(1 2 3 7)</td>
<td>11</td>
<td>18</td>
<td>25</td>
<td>32</td>
<td>39</td>
<td>46</td>
</tr>
<tr>
<td>5</td>
<td>(1 2 4 5)</td>
<td>12</td>
<td>19</td>
<td>26</td>
<td>33</td>
<td>40</td>
<td>47</td>
</tr>
<tr>
<td>6</td>
<td>(1 2 4 6)</td>
<td>13</td>
<td>20</td>
<td>27</td>
<td>34</td>
<td>41</td>
<td>48</td>
</tr>
<tr>
<td>7</td>
<td>(1 2 4 7)</td>
<td>14</td>
<td>21</td>
<td>28</td>
<td>35</td>
<td>42</td>
<td>49</td>
</tr>
</tbody>
</table>

According to the property of combination, let \( \alpha = (a_1, a_2, \ldots, a_m) \in C(n, m) \) and \( a_i = j \). Then the number of combinations in \( C(n, m) \) as \( C_{m-j}^{n-i} \) by Proposition 3. Thus we can get a matrix \( M(T(i, j))_{m\times n} \) about \( C(n, m) \), where

\[
T(i, j) = \begin{cases} C_{m-j}^{n-i} & 1 \leq i \leq m, i \leq j \leq n + 1 - m \\ 0 & \text{other.} \end{cases}
\]

The follow equation is apparent: if \( T(i, j) \neq 0 \), then

\[
T(i, j) = \sum_{k=j}^{n} T(i+1, k).
\]

**Example 4.** For \( C(7, 4) \), matrix \( M(T(i, j))_{7\times 4} \) is as follows:

\[
M = \begin{pmatrix} 20 & 10 & 4 & 1 & 0 & 0 & 0 \\ 0 & 10 & 6 & 3 & 1 & 0 & 0 \\ 0 & 0 & 4 & 3 & 2 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}.
\]

Since \(|\{1, 2, \ldots, C_n^m\}| = |C(n, m)| = C_n^m \), define a mapping \( h \) from \( \{1, 2, \ldots, C_n^m\} \) to \( C(n, m) \) as follows:

\[
h : \{1, 2, \ldots, C_n^m\} \rightarrow C(n, m)
\]

\[
h(x) = \alpha = (a_1, a_2, \ldots, a_m),
\]

where \( x \in \{1, 2, \ldots, C_n^m\} \) and \( \alpha = (a_1, a_2, \ldots, a_m) \in C(n, m) \). We will give Algorithm 1 for calculating the combination by the matrix \( M(T(i, j))_{m\times n} \).

For all \( x \in \{1, 2, \ldots, C_n^m\} \), according to Algorithm 1 and (9) and (10), we can calculate \( h(x) = \alpha = (a_1, a_2, \ldots, a_m) \); thus we can decode a integer within \( C_n^m \) into a combination in \( C(n, m) \).

**Example 5.** \( C_7^4 = 35; \) let \( x = 22 \in \{1, 2, \ldots, 35\} \) and \( h(22) = (a_1, a_2, a_3, a_4) \). By Algorithm 1, we know

\[
20 = T(1, 1) < x = 22 < T(1, 1) + T(1, 2) = 30; \text{then } a_1 = 2, x = 22 - T(1, 1) = 2;
\]

\[
x = 2 < T(2, 2+1) = 6; \text{then } a_2 = 3, x = 2 - 0 = 2;
\]

\[
x = 2 < T(3, 3+1) = 3; \text{then } a_3 = 4, x = 2 - 0 = 2;
\]

\[
1 = T(4, 4+1) < x = 2 < T(4, 4+1) + T(4, 4+2) = 2; \text{then } a_4 = 6;
\]

\[
h(22) = (2, 3, 4, 6);
\]

in Table 1, the 22th permutation being \( (2, 3, 4, 6) \), that is, the same result.
input: x, n, m. (x ≤ Cⁿₘ)  
output: α = (a₁, a₂, ..., aₘ)  
begi
    ℓ = 1;
    if ∑_{k=1}^{ℓ} T(1, k) < x ≤ ∑_{k=1}^{ℓ−1} T(1, k);
    a₁ = j₁ and x = x − ∑_{k=1}^{ℓ−1} T(1, k);
    ℓ = ℓ + 1;
    end
for i = 2 to m do
    if ∑_{k=1}^{i} T(1, k) < x ≤ ∑_{j=1}^{i−1} T(1, j);
    aᵢ = jᵢ, and x = x − ∑_{j=1}^{i−1} T(1, j);
end
end

Algorithm 1: DetCoCo(x, n, m).

4. RFSA-RST Algorithm

In this section, for finding minimal reducts of a dataset, an algorithm which is named RFSA-RST in this paper based on RST and FSA is proposed. It first uses the concepts of the core to find the core attributes, and FSA is employed in a restrained manner to find the minimal reducts to ensure that the result is the minimal length.

4.1. Encoding Method. Essentially, AR is a combinatorial optimization problem and the solution space is the power set of the attribute set. Let C = {1, 2, ..., n} be conditional attribute set, the power set of C be 2ⁿ, and C(n, m) = {α | α ⊆ C, |α| = m}; then C(n, m) ⊆ 2ⁿ. Clearly, ∑_{i=0}^{n} C(n, i) = C and C(n, i) ∩ C(n, j) = φ (i ≠ j). By the cardinality of the subset of C, 2ⁿ is split into n disjoint parts. According to Algorithm 1, every integer within Cₘⁿ can be decoded into a combination in C(n, m), so integer encoding method will be adopted. Thus, each integer represents a combination in C(n, m) and the solution space is integer type. Let ∆ be defined as the increment of Xᵢ. By (4)–(6), it may be decimal. So, it is not adaptive to the solution space. In order to improve this problem, the new Xᵢ increment is shown as follows [11]:

\[
Xᵢ + \text{round}(\Delta) + 1 \quad \Delta > 0
\]

\[
Xᵢ + \text{round}(\Delta) - 1 \quad \text{else}
\]

4.2. Fitness Function. In FSA, the quality of a AF is determined by fitness function. For AR problem, the subset should retain the classification ability and has minimal attributes. Generally, the fitness function must meet those issues. In addition, it should be as simple as possible to improve the computation efficiency. In this paper, all subsets without containing the core attribute are encoded into an integer as a AF. Since the cardinality of each combination in C(n, m) is fixed, FSA is restricted to a fixed length subset of condition attribute set and judges a AF by its dependency value for the attribute subset represented by the integer, as explained below.

4.3. Algorithm Description. In FSA, the initial population of size N needs to be provided representing multiple points in the search space. If there is a fitness value of AF that is equal to γₜ, then the process terminates and outputs the corresponding attribute subsets of AF. If change in average fitness value of two successive generations is 0.001 or less, no further generations are created and terminate this loop. Then a new initial population is created for the next loop. These attribute subsets of the new initial population correspond to one more attribute. So, initial strings in ith loop of the algorithm have one more attribute than that in (i − 1)th loop. The whole process is repeated based on this new initial population until a minimal set of reduct(s) is found. The detailed process of RFSA-RST is shown in Algorithm 2.

5. Performance Comparison

To evaluate the effectiveness of the proposed algorithm, RFSA-RST, we carried out experiments in the following way. Another two different types of algorithms were used. One is proposed in [12], denoted here by RGA-RST, and the other is proposed in [17], denoted here by PSO-RST. A PC running Windows XP with 3.1 GHz CPU and 1 GB of main memory was employed to run these three algorithms. The test datasets were obtained from the UCI machine learning repository and 6 datasets were chosen. Basic information about the datasets is listed in Table 2, where n and m are the number of objects and (condition) attributes, respectively.

![Table 2: Basic information about the datasets.](image)

For a AF and its position being X, calculate the corresponding combination h(x) in C(n,m) by Algorithm 1. Define the fitness value of X as follows:

\[
F(X) = γ_{h(X)∪Core}^C
\]
Input: $DS = (U,C,D,V,f)$.
Output: $Red$

begin
  $Red = \{}$; Calculate $\gamma_C$ and $\gamma_{Core}$;
  if $\gamma_{Core} \neq \gamma_C$,
    $Red = Core$, output $Red$ and terminate process;
  else
    $L = C \setminus Core$; $n = |L|$; $m = 0$;
    repeat
      $m = m + 1$; $avg_o = 0$;
      Get $\{r_1, r_2, \ldots, r_N\} \subseteq \{1, 2, \ldots, C^n\}$ randomly;
      repeat
        $sum = 0$;
        for $i = 1$ to $N$
          Calculate $P_i = f(r_i)$ by Algorithm 1; $P_i = P_i \cup CS$;
          Calculate $F(P_i)$ for the permutation $P_i$, by Eq (14).
          if $F(n_i) = \gamma_C$
            $Red = Red \cup P_i$
          else
            $sum = sum + F(P_i)$;
          end
        end
        $avg_n = sum/N$; $change = avgn - avg_o$; $avg_o = avgn$;
        Implement swarming behavior, following behavior
        and preying behavior.
        until $change \leq 0.001$;
      until $Red \neq \{}$;
      return $Red$;
    end
end

Algorithm 2: RFSA-RST(DS).

Table 3: Performance of three algorithms.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>PSO-RST</th>
<th>RGA</th>
<th>RFSA-RST</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>STL</td>
<td>AVL</td>
<td>STL</td>
</tr>
<tr>
<td>Lung-cancer</td>
<td>3</td>
<td>4.14</td>
<td>3</td>
</tr>
<tr>
<td>Soybean-small</td>
<td>2</td>
<td>2.29</td>
<td>2</td>
</tr>
<tr>
<td>Sponge</td>
<td>8</td>
<td>8.7</td>
<td>8</td>
</tr>
<tr>
<td>Zoo</td>
<td>5</td>
<td>5.32</td>
<td>5</td>
</tr>
<tr>
<td>Blance</td>
<td>4</td>
<td>4.02</td>
<td>4</td>
</tr>
<tr>
<td>Spect</td>
<td>16</td>
<td>16</td>
<td>16</td>
</tr>
</tbody>
</table>

algorithms on the datasets are reported in Tables 3 and 4, respectively.

From Table 3, the proposed algorithm has the same performance as the other two algorithms for the shortest length of the outputs, but outperforms the other two algorithms in terms of the average length of output solutions except for the Soybean-small dataset. It means that the stability of the proposed algorithm is higher than the other two algorithms.

From Table 4, for the ratio of normal runs, all the outputs of the three algorithms are roughly the same, but the proposed algorithm outperforms the other two algorithms in terms of the ratio of successful runs except for the Soybean-small dataset. Therefore, if the minimum attribute reduction is required, the proposed algorithm is the best of the three algorithms. It is also reflected to the average length in Table 3. As far as the average running time is concerned, the proposed algorithm is slightly worse than PSO-RST but better than RGA.

Since the operation modes of the proposed algorithm and RGA algorithm are the same, another experiment has been done to evaluate the time efficiency of the two algorithms. We modify some process of both algorithms where the core attribute set is no longer identified and $n$ in RGA (for the parameter $m$ in the proposed algorithm) is set from 1 to the cardinality of minimal attribute reduction for each dataset. Record the running time for each loop. In order to show the results more clearly and length limitations, we only show the test results about the datasets Spect and Sponge. As shown
Table 4: Performance of three algorithms.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>PSO-RST $s_1/s_2$</th>
<th>AVT(s)</th>
<th>RGA $s_1/s_2$</th>
<th>AVT(s)</th>
<th>RFSA-RST $s_1/s_2$</th>
<th>AVT(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung-cancer</td>
<td>0.06/1</td>
<td>0.2073</td>
<td>0.08/1</td>
<td>0.8103</td>
<td><strong>0.22/1</strong></td>
<td>0.2255</td>
</tr>
<tr>
<td>Soybean-small</td>
<td>0.7/1</td>
<td>0.7117</td>
<td>0.62/0.98</td>
<td>16.1129</td>
<td><strong>0.78/0.98</strong></td>
<td>6.8089</td>
</tr>
<tr>
<td>Sponge</td>
<td>0.6/0.92</td>
<td>6.4159</td>
<td>1.3892</td>
<td>2.1648</td>
<td>1/1</td>
<td>1.4584</td>
</tr>
<tr>
<td>Zoo</td>
<td>0.94/1</td>
<td>0.5726</td>
<td>1/1</td>
<td>1.9935</td>
<td>1/1</td>
<td>0.6045</td>
</tr>
<tr>
<td>Blance</td>
<td>0.98/1</td>
<td>12.0372</td>
<td>36.0937</td>
<td>12.9012</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spect</td>
<td>1/1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

in Figure 1. By Figure 1, we can find that, for all datasets, the running time in each loop of the proposed algorithm was increased with the increase of $m$, while RGA were relatively stable. That is due to the fact that the search process of the proposed algorithm focuses on the attribute subset with fixed length. Therefore, the time efficiency of the algorithm is higher in each loop and the convergence rate is faster. It shows that the algorithm proposed in this paper is more efficient.

Summing up the experiment results, we see that the proposed algorithm is more efficient than the other two typical algorithms on the datasets for test, although the running time is slightly worse than PSO-RST.

6. Conclusions

In this paper, we derived a new method for minimum attribute reduction based on rough set theory and fish swarm algorithm. In this method, by a coding method of combination subset of attributes set, the FSA has been used to search minimum attribute reduction and attribute dependency has been applied to calculate the fitness values of the attribute subsets. The FSA has been restrained in which every integer corresponds to a specified length attribute subset in each search process. The cardinality of the attribute subset represented by AF was starting from the length of the core and incremented by one after each loop.

Numerous test results show that it can improve not only the accuracy of minimal attribute reduction but also the efficiency of convergence rate.

To improve search efficiency, future enhancements to this work are to confirm whether the starting length was reasonable and, to improve the search time efficiency, these works can reduce redundant search process.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work is supported by the Opening Project of Sichuan Province University Key Laboratory of Bridge Non-Destruction Detecting and Engineering Computing (2014QYJ02).

References


Submit your manuscripts at https://www.hindawi.com